Neural Mechanisms for Face and Orientation Aftereffects

Chen Zhao

Doctor of Philosophy

Institute for Adaptive and Neural Computation

School of Informatics

University of Edinburgh

2011

Abstract

Understanding how human and animal visual systems work is an important and still largely unsolved problem. The neural mechanisms for low-level visual processing have been studied in detail, focusing on early visual areas. Much less is known about the neural basis of high-level perception, particularly in humans. An important issue is whether and how lessons learned from low-level studies, such as how neurons in the primary visual cortex respond to oriented edges, can be applied to understanding high-level perception, such as human processing of faces. Visual aftereffects are a useful tool for investigating how stimuli are represented, because they reveal aspects of the underlying neural organisation. This thesis focuses on identifying neural mechanisms involved in high-level visual processing, by studying the relationship between low- and high-level visual aftereffects.

Previous psychophysical studies have shown that humans exhibit reliable orientation (tilt) aftereffects, wherein prolonged exposure to an oriented visual pattern systematically biases perception of other orientations. Humans also show face identity aftereffects, wherein prolonged exposure to one face systematically biases perception of other faces. Despite these apparent similarities, previous studies have argued that the two effects reflect different mechanisms, in part because tilt aftereffects show a characteristic S-shaped curve, with the effect magnitude increasing and then decreasing with orientation difference, while face aftereffects appeared to increase monotonically (in various units of face morphing strengths) with difference from a norm (average) face. Using computational models of orientation and face processing in the visual cortex, I show that the same computational mechanisms derived from early cortical processing, applied to either orientation-selective or face-selective neurons, are sufficient to replicate both types of effects. However, the models predict that face aftereffects would also be S-shaped, if tested on a sufficiently wide range of face stimuli.

Based on the modelling work, I designed psychophysical experiments to test this theory. An identical experimental paradigm was used to test both face gender and

tilt aftereffects, with strikingly similar S-shape curves obtained for both conditions. Combined with the modelling results, this result provides evidence that low- and high-level visual adaptation reflect similar neural mechanisms.

Other psychophysical experiments have recently shown interactions between low- and high-level aftereffects, whereby orientation and line curvature processing (in early visual area) can influence judgements of facial emotion (by high-level face-selective neurons). An extended multi-level version of the face processing model replicates this interaction across levels, but again predicts that the cross-level effects will show similar S-shaped aftereffect curves. Future psychophysical experiments can test these predictions.

Together, these results help us to understand how stimuli are represented and processed at each level of the visual cortex. They suggest that similar adaptation mechanisms may underlie both high-level and low-level visual processing, which would allow us to apply much of what we know from low-level studies to help understand high-level processing.

Acknowledgements

I would like to first show my gratitude to my supervisor Jim Bednar. I would not have been able to get through my academic life and finish this thesis in the past five years without his invaluable support, encouragement and inspiration.

I would also thank my second supervisor Peter Hancock for his helpful advice on psychophysical experiments and important data on face processing, as well as my third supervisor Peggy Seriés for her advice on modelling. I would like to acknowledge DTC and Institute for Adaptive and Neural Computation for their financial and administrative support which makes my academic pursuits possible.

Thanks to my fellow PhD students and lab members for their kind help and advice. Especially, I would like to thank Jan Antolik, Chris Ball, Judith Law and Chris Palmer for their contributions of time and energy on helping me finish my psychophysical experiments. My thanks are also to other experiment participants Liang Wang and Zhiguo Kang. The achievements described in this thesis cannot be possible without their help.

Finally and most importantly, I would like to thank my dearest mother and father who have given me enormous emotional and financial support throughout the past five years. It is their countless phone calls and emails that kept me pursuing in academia faraway from homeland.

Chen (Roger) Zhao

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Chen Zhao*)

Table of Contents

List of Figures

Chapter 1

Introduction

Starting from the retina, optical information is turned into electrical signals, flows among various cortical areas, and somehow forms the perceptual experience of "vision". The final perception involves the activation and subsequent adaptation of trillions of interconnected neurons in many visual cortical areas. The adaptation process usually goes unnoticed, but in some special cases results in changes in how accurately we perceive the world, known as "visual aftereffects". Understanding the neural mechanisms underlying visual aftereffects can greatly help to elucidate the general mechanisms governing visual processing, which could finally answer the question "How can we see?"

Over the past two hundred years, many types of aftereffects have been found, from the famous "Waterfall Illusion" to the more recently discovered aftereffects of face perception (for a review, see Thompson and Burr, 2009). These aftereffects are generally induced by a particular kind of visual pattern, which range from simple geometric patterns (such as the oriented lines leading to a tilt aftereffect, TAE) to complex visual patterns or objects (such as faces, leading to a face aftereffect, FAE).

Research in experimental neuroscience on animals and humans suggests that visually responsive neurons are connected into a rough hierarchy of cortical areas. In the classic studies by Hubel and Wiesel (1968), neurons in a low-level cortical area such

as the primary visual cortex (V1) of monkeys were found to have visual responses that were selective to the orientation of a line or an edge. More recent imaging studies on humans suggest that some cortical regions, such as the fusiform face area (FFA), respond selectively to photographs of human faces (Sergent et al., 1992; Kanwisher et al., 1997).

As will be described in detail in this thesis, aftereffects for simple patterns as in the TAE can be explained in terms of low-level mechanisms operating in V1. However, the neural substrate for aftereffects for more complex stimuli like faces is not yet known. A conservative hypothesis is that high-level visual processing is based on the integration of feedforward information from V1 and other lower cortical levels (e.g., Tanaka, 1996; Logothetis and Sheinberg, 1996; Perrett and Oram, 1993).

Understanding how similar such integration is to the processing in lower-level areas like V1 is crucial for determining how much of our knowledge of V1 can be applied to higher visual areas like the FFA or to the cortex as a whole. High-level effects like the FAE could provide important clues for this overall project of understanding visual and other cortical processing.

1.1 Problem and contributions

This thesis focuses on identifying the properties of the neural mechanisms involved in high-level visual processing, by studying the relationship between low- and high-level visual aftereffects. This thesis addresses one main issue — can (and do?) high-level aftereffects arise from the same types of mechanisms that lead to low-level aftereffects? To address this issue we consider the following questions:

1. Computationally, how can face aftereffects arise from general mechanisms of neural processing?

2. Computationally, can face aftereffects arise from mechanisms previously demonstrated to lead to low-level effects like tilt aftereffects?

3. Psychophysically, is there any evidence to support this similarity?

4. What computational effects will result from low-level and face aftereffects propagating from low-level to high-level cortical areas?

In this thesis, the tilt aftereffect (TAE) is the main example of a low-level visual aftereffect, and the facial gender aftereffect (FAE) is the main example of a high-level visual aftereffects. Both of these were chosen because they are the best-studied examples of low-level and high-level patterns known to lead to aftereffects. The TAE has been studied extensively over the past 75 years, while FAEs and face processing in general have received intensive scrutiny more recently.

By combining computational modelling and psychophysical experiments, this thesis presents answers to the above four questions, in order to provide an experimental and computational account for the relationship between low-level and high-level aftereffects. The computational results show how existing psychophysical findings can be replicated using computational models, and make novel predictions about the aftereffects that are testable using psychophysical experiments. Some of these predictions were tested experimentally, verifying the predictions of the model and thus strongly suggesting that similar underlying mechanisms lead to both types of effects (contrary to much of the previous literature). This thesis should thus offer important constraints on and guidance for future investigation of the neural representation of complex visual stimuli, eventually helping to gain an understanding of how the human visual system works.

1.2 Thesis structure and rationale

Chapter 2 reviews the previous work on the TAE and FAE, including face space theory, physiological and psychophysical experiments on low-level aftereffects and face aftereffects, and previous models of face and object processing. This chapter introduces all the relevant background that the work of this thesis is based on.

Chapter 3 addresses how face aftereffects can arise and be represented in a computational model. Starting by duplicating the face identity aftereffect found by Leopold et al. (2001), this chapter shows how this FAE can arise from the same underlying mechanisms that produce TAEs, and illustrates how the FAE can be represented as shifts of perception.

Chapter 4 builds on modelling results from Chapter 3, reducing the multidimensional FAE to a simple one-dimensional form that can be compared more directly with low-level aftereffects in psychology. This chapter shows that both the general modelling approach from Chapter 3, and a simpler model tailored specifically to one-dimensional aftereffects, predict FAEs and TAEs with similar curves.

Chapter 5 describes psychophysical experiments designed explicitly to test the predictions from Chapter 4. A simple paradigm is used to test face gender aftereffects on human participants. The identical paradigm is also applied to the TAE. Results show that both effects exhibit similar S-shaped aftereffect curves, as was predicted from the models based on low-level processing as in the TAE, but contrary to predictions from other previously proposed models.

Chapter 6 addresses a further question: if an aftereffect at a low level changes the neural activity patterns that propagate to higher cortical levels, how will high-level perception be affected? This chapter replicates an important part of the experimental work conducted by Xu et al. (2008) in a computational model. The model first illustrates how such activity propagation could operate, replicating the previous psychophysical results, and then predicts that the propagated adaptation will also exhibit an S-shaped aftereffect curve.

Chapter 7 discusses the implications of the work in this thesis and suggests areas for future study.

Chapter 2

Background and related work

This chapter reviews the relevant background for this thesis, including face space theory, physiological and psychophysical experiments on low-level aftereffects and face aftereffects, and previous models of face and object processing.

2.1 Physiology of face processing

Visual object recognition is at the centre of understanding how humans see. This amazing capability is crucial for human beings to interact with and reason about the world. This capability is thought to be achieved through a roughly hierarchical visual system. In this hierarchy, starting from the retina, neurons in low-level areas such as the lateral geniculate nucleus (LGN) and primary visual cortex (V1) respond selectively to specific simple visual patterns like lines and edges. Complex patterns consisting of combinations of these elements are processed further by neurons in the high-level cortical areas such as inferior temporal cortex (IT) and the fusiform face area (FFA). Along the hierarchy, neurons become selective for more complex visual patterns (Kobatake and Tanaka, 1994).

An influential hypothesis is that this system is organised into two distinct pathways. As visual information exits the early visual areas, it follows two main "streams": the dorsal stream and the ventral stream (Mishkin and Ungerleider, 1982; reviewed in

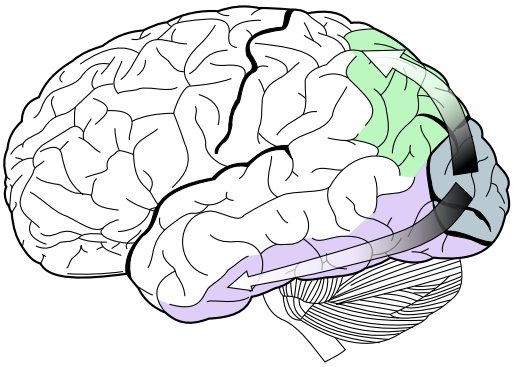


Figure 2.1: The dorsal stream (green) and ventral stream (purple) are shown. They originate from a common source in the visual cortex. Source: `http://en.wikipedia.org/wiki/File:Ventral-dorsal_streams.svg`

Goodale and Milner, 1992). Figure 2.1 shows a schematic plot of the flow of these two streams. The dorsal stream (or "where" stream) is mainly involved in spatial attention, global patterns of motion, and eye movements, and guides behaviour to spatial locations. The ventral stream (or "what" stream) is involved in the recognition, identification and categorisation of specific visual stimuli. The proposed specialisation of visual functions has been controversial as the two streams are effectively heavily interconnected (e.g., Farivar et al., 2009). In any case, it is clear that the high-level regions in the "what" stream are heavily involved in object identification tasks, which are the focus of this thesis.

Among the many cognitive functions of the high-level visual system, face recognition is an important one, as it is crucial for identifying individuals and interpreting facial expressions and eye gaze during social interaction. For this reason, the human face is the most commonly studied type of complex visual stimulus. These studies date to findings by Gross et al. (1969) and Gross et al. (1972) that neurons in the IT cortex of macaques responded most strongly to complex visual stimuli, such as hands and faces.

Gross (1994) and Rodman (1994) reviewed studies of IT as a highly organised structure where single units code objects. Even a single neuron — the famous grand-

mother cell — has been claimed to be tuned to a particular object (Konorski, 1967). Tanaka (1996) found a highly structured pattern of feature selectivity by recording within the temporal area E (TE) of IT. However, most researchers argue that individual neurons are not sufficient for discriminating between all the possible objects. Instead, populations of neurons can be used to represent objects more plausibly (Logothetis and Pauls, 1995; Booth and Rolls, 1998; Perrett et al., 1998; Kobatake et al., 1998).

In terms of how neurons encode faces, Haxby (2006) suggested that individual neurons within a highly face-selective cluster are differentially but coarsely tuned to individual faces. Activity patterns over a population of such neurons can thus encode face identity. Treisman (1996) suggested that sparseness can help simultaneous encoding of more than one stimulus. Gilaie-Dotan and Malach (2007) showed that narrow neural tuning coincides with high sensitivity to face changes. Together, these results suggest that an individual face is processed by a subset of the neurons in face-selective areas, and that different faces will activate different subsets.

How specialised this processing is for faces is still controversial. It is widely accepted that face processing is to some degree "special" compared to other object processing, but how exactly such a specialisation has arisen and whether there are inherently face-dedicated cortical regions is still under debate.

In terms of the speciality of face processing, studies on infants showed that newborn babies have the ability to discriminate a face-like stimuli from a non-face stimuli, suggesting that face processing may be genetically encoded (Johnson and Morton, 1991; Mondloch et al., 1999). Many authors thus argue that faces are processed qualitatively differently than other visual stimuli (Farah et al., 1998; McKone et al., 2007). Sergent et al. (1992) and later Kanwisher et al. (1997) proposed a neural substrate for these differences, namely that neurons in a fusiform face area (FFA) on the ventral surface of the temporal lobe are all highly selective to face stimuli, and thus FFA might be specialised for face recognition.

However, other evidence does not support the idea of "natural encoding" and ded-

icated organisation. The series of work by Gauthier and colleagues showed that human participants could be trained as experts in an artificial non-face domain of "greebles", developing capabilities competitive with those for recognising faces (Gauthier and Tarr, 1997; Rossion et al., 2004; Bukach et al., 2006). Further studies also showed that expertise in "greebles" increases specific neural selectivity in the FFA (Gauthier et al., 1999). A recent high-resolution fMRI study on the fusiform face area (FFA) also found highly selective non-face clusters (Grill-Spector et al., 2006). Previously, it had been thought that the FFA was composed of only face-selective neurons (Kanwisher et al., 1997), and some authors have found other areas full of face-selective neurons (Tsao et al., 2006). The current view is that the FFA is a heterogeneous region composed of highly selective neural populations for both faces and non-faces, mixed together, but with a larger number of neurons that are selective to faces than to non-faces.

In any case, it is widely accepted that even if face processing is unique and special, it can still be greatly rewired by experience (for a review, see Park et al., 2009). Other physiological details are remain largely unknown, for example, how the face is encoded and decoded, how the speciality of face processing arises, and how it is enhanced by experience. A serious barrier to progress in this area is that most face studies require human participants, and are thus limited to non-invasive forms of testing, unlike corresponding studies of simple patterns processed in V1.

2.2 Psychophysics of face aftereffects

Psychophysical studies try to address the problem of face processing through behavioral experiments, typically by investigating the results of visual adaptation. Visual adaptation is change in responsiveness or sensitivity resulting from prolonged exposure to a particular type of visual stimulus. Both perception and the response properties of neurons to subsequent stimuli are affected by sensory stimuli. Adaptation

typically applies to the sensory experience of the preceding tens of milliseconds to minutes. Adaptation can cause a particularly striking and easily studied form of plasticity that can be measured as aftereffects — the misjudgement of cognitive perception. This way, investigation into the aftereffects caused by face stimuli can help to gain an understanding of the properties of sensory system change, without requiring invasive measurements of the underlying neural substrates.

In recent years, psychophysicists have found many kinds of face aftereffects, including face identity (Leopold et al., 2001), gender (Webster et al., 2004), distortion (Robbins et al., 2007), emotion (Xu et al., 2008), viewpoint (Chen et al., 2010), illumination (Oruç and Barton, 2010), etc.

The first face aftereffect was found by Leopold et al. (2001). They showed that humans will systematically misjudge face identity, due to their recent experience with other faces. The misjudgements appear to be related to the location of each face along various dimensions that together represent a multidimensional "face space".

The stimulus examples, experimental paradigm and results of this experiment are shown in Figure 2.2. The face stimuli used in the experiment were drawn from a multidimensional space where each trajectory (shown as a grey line in Figure 2.2a) represents a particular face dimension. In the work of Leopold et al. (2001), the face space was generated by computing an average face (blue ellipse), then morphing the original faces (shown as green ellipses in the figure), to generate anti faces (red ellipses) and other faces along this trajectory. The generated face trajectory can be indexed by the value of "identity strength". The experimental paradigm was composed of three adaptation conditions: baseline, match and non-match (Figure 2.2b) trials. In the baseline trials, the participants were asked to judge the original identity of a 0.25 identity strength face. In the match and non-match trials, the participants were first adapted to the anti-face at the same trajectory or at a different trajectory before the judgement. It can be seen from the results shown in Figure 2.2c that the judgement performance was improved in the match trials (closed circles), but impaired in the non-match trials

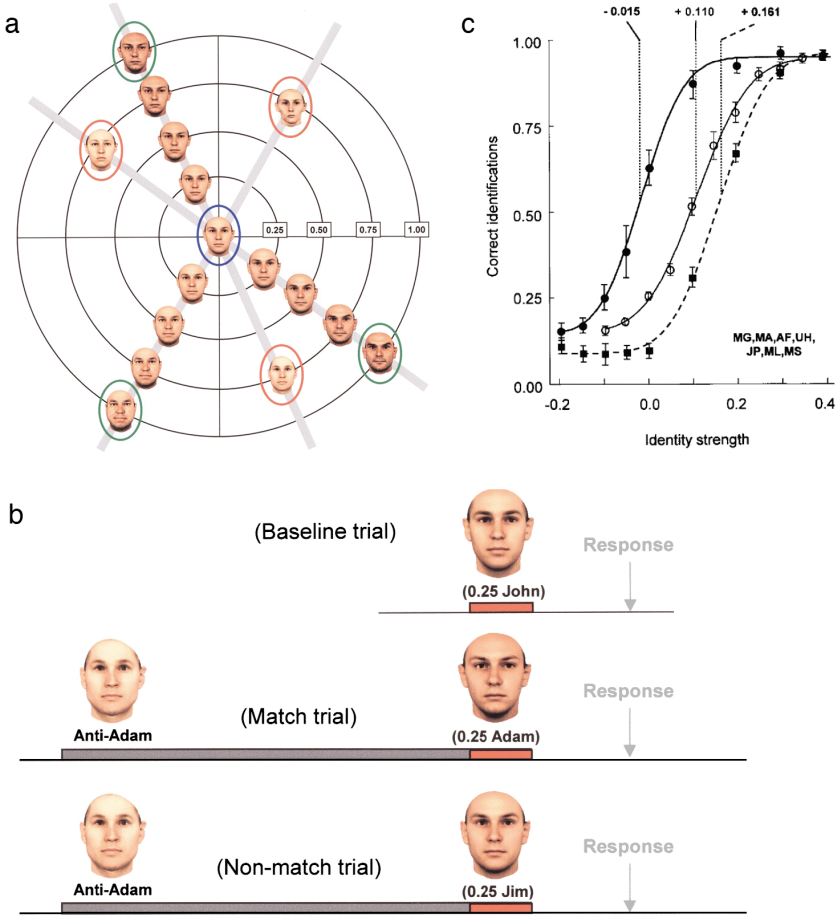


Figure 2.2: Leopold et al. (2001) experiments. **(a)**: Examples of the stimuli and the face space. The original faces (green ellipses) are connected to the average face (blue ellipse) by an identity trajectory. The position of a face along its identity trajectory can be manipulated and thus its level of distinctiveness or identity strength can be systematically varied. Such identity strength (shown numerically in the figure) refers to the identity strength possessed by the given face in the dimension. (For example, the veridical face equals 1.0, and the average prototype face equals 0.0 in the centre of the face space.) **(b)**: Experimental paradigm. Top row: baseline trial without adaptation where the participant simply selects the current original (identity strength = 1.0) face matching the given face; mid row: match trial where the participant is first adapted to anti-Adam before choosing which of four given faces are most like 0.25 Adam; bottom row: non-match trial where the participant is first adapted to anti-Adam before choosing which of the four faces are most like 0.25 Jim (another identity). In each trial, the presentation duration is 5 seconds for the adaptation faces and 0.2 seconds for the test stimuli. **(c)**: Results for the average human performance under the conditions: no adaptation (open circle), adapting to matching anti-face (closed circle) and adapting to non-matching anti-face (closed square). Adaptation thus systematically shifts the perception of original (identity strength = 1.0) faces.

(closed square).

While the work by Leopold et al. (2001) tested the facilitation of adaptation in face identity judgement, the experiments by Webster et al. (2004) tested the perceptual boundary shifts following adaptation to face gender, ethnicity and expression. The experimental procedure of their work was similar to that of Leopold et al. (2001), but instead of making a judgment of original face identity, the participant was asked to answer the question of gender (male or female), ethnicity (Caucasian or Japanese) or expression (e.g., happy or angry) as a binary choice. Their experiments were based on the face trajectories of gender, ethnicity and expression. As shown in Figure 2.3, it can be seen that in most cases, the directions of boundary shift are identical to the directions of the adaptation condition. For example, in Figure 2.3a, the perceptual boundary shifts towards a male direction when adapting to a male face, compared to the position in the neutral condition.

The Webster et al. (2004) result is qualitatively similar to the experiments conducted by Robbins et al. (2007) on the distorted face continuum, i.e., the perceptual boundary of a normal face shifts towards the direction of compacted faces when adapting to a compacted face.

Psychophysical experiments on face aftereffects in recent years have begun to pay attention to the role of adaptation to low-level elements in face adaptation. Xu et al. (2008) showed that adaptation to simple low-level stimuli such as curved lines can affect emotion judgement for a cartoon face or a photographic face. An example of their paradigm is shown in 2.4. The experimental procedure was similar to those used by Webster et al. (2004) and Robbins et al. (2007). Samples of real face stimuli and results are shown in Figure 2.5. It can be seen that adaptation to the curved line stimulus can induce a boundary shift (red line) although the magnitude is not as big as the photographic face (green line). Note that the trials in Figure 2.5b are both adaptation to the strength 0 face (the scale in the figure), and so the shift direction is also consistent with those in the face gender (Webster et al., 2004) and distortion (Robbins et al.,

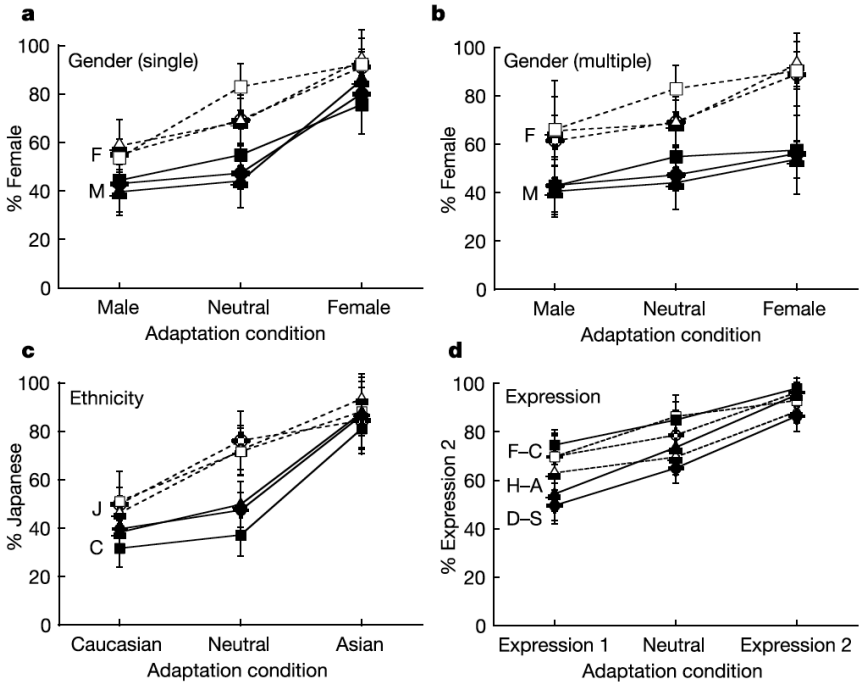


Figure 2.3: Category boundaries before or after adaptation. Each plot shows the mean boundaries set by two observers for three face pairs, before or after viewing images from the two categories defining the morphs. On adaptation trials, the initial adaptation time is 180 seconds, and then 5 seconds in each trial; the presentation duration for the test stimulus is 0.5 seconds. **(a)**: Settings by a female (F) or male (M) for gender morphs. **(b)**: Settings by the same observers after adapting to a series of different male or female faces. **(c)**: Settings by a Japanese (J) or Caucasian (C) observer for ethnicity morphs. **(d)**: Boundaries selected by a Caucasian female (open symbols) and male (filled symbols) for morphs between happy-angry (H-A, triangles), disgust-surprise (D-S, circles) or fear-contempt (F-C, squares). For most cases, the PSE (50% female point) on each curve moves towards the direction of adaptation. Source: Webster et al. (2004)

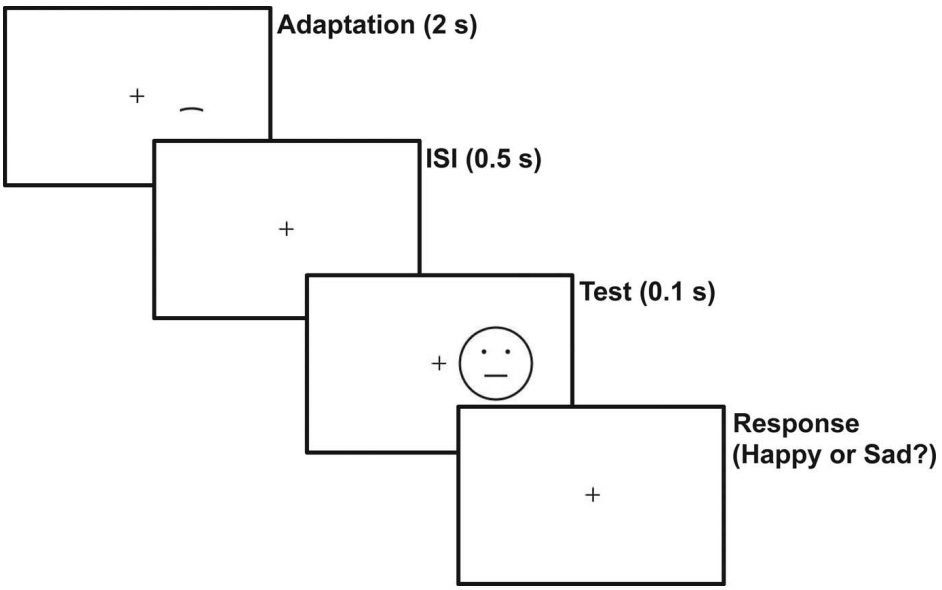


Figure 2.4: 2AFC paradigm of the psychophysical experiments conducted by Xu et al. (2008). This diagram shows the process of adaptation to a mouth curve and testing on a cartoon face. The procedure used for the other experiments described in this work is the same apart from the different adaptation or test stimuli. Either a cartoon face (as shown) or real face can be used. Source: Xu et al. (2008).

2007) aftereffects. However it is not clear from the study which magnitude is bigger (adaptation to face, then on curve, and adaptation to curve, then test on face), as the participants in these two conditions are different. This work shows that the low-level induced adaptation can produce qualitatively similar aftereffects as high-level induced adaptation.

Another type of work involves different types of elements such as viewpoint or contrast. Chen et al. (2010) tested face viewpoint aftereffects produced by photographic faces at different viewpoints. Oruç and Barton (2010) tested face contrast aftereffects produced by photographic faces at different illumination levels. Again, the results of their perceptual boundary shift are both consistent with the experiments above.

A recent study has looked for physiological correlates for the psychological effects measured in FAE experiments. Leopold et al. (2006) showed that the face-responsive neurons in the macaque anterior IT cortex were tuned to a similar face space (to be

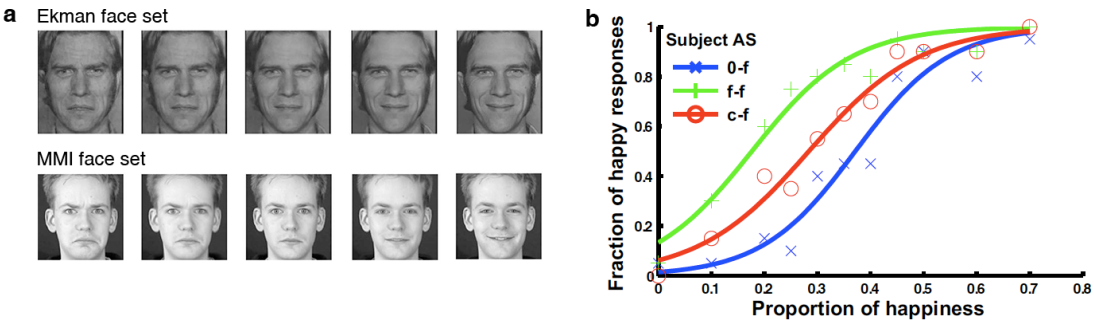


Figure 2.5: Examples of real face stimuli and the effect of curve adaptation on the perceived expression of the real faces, both from the experiments conducted by Xu et al. (2008). **a**: The Ekman face set (Ekman and Friesen, 1976) and the MMI face set (Pantic et al., 2005) were both used in the human psychophysical experiment. **b**: Psychometric functions from a naive subject for the perception of the Ekman faces under the following conditions. 0-f, no adaptation baseline (blue); f-f, adaptation to the saddest face (green); c-f, adaptation to the concave curve, whose curvature and length matched those of the saddest face (red). For each condition, the fraction of happy responses was plotted as a function of the proportion of happiness in the morphed test faces. A proportion of 0 or 1 corresponded to the original sad or happy image taken from the Ekman database. Source: Xu et al. (2008)

introduced in Section 2.4.1) found in human psychophysical experiments. The norm face had the lowest response, and the face-selective responses were shaped by the contrast between the test face and the face norm. This evidence then links the underlying tuning of neurons and the psychophysical aspects, suggesting a potential theory combining low- and high-level processing.

2.3 Psychophysics of low-level aftereffects

Compared to high-level face aftereffects, low-level aftereffects, usually induced by simple visual patterns, have been known for hundreds of years and have been studied in much more detail than high-level face aftereffects. The most well-studied low-level aftereffect is the tilt aftereffect (TAE) found by Gibson and Radner (1937). After prolonged exposure to a pattern of tilted lines or gratings, the subsequent lines appear to have a slight tilt in the opposite direction. Following the instructions in Figure 2.6a, the TAE should appear for most participants. If more adaptation conditions (gratings at different orientations) are tested, a resulting aftereffect curve, like the one in Figure 2.6b, can be obtained. It can be seen that for the adaptation conditions immediately near degree 0, the boundary shift directions are also consistent with the results for the face aftereffects reviewed in Section 2.2. But the TAE shows an S-shape curve for more distant patterns, while the face aftereffects do not in the studies so far.

Note that when referring to aftereffects for simple geometric stimuli as “low-level”, it is implicitly assumed that the mechanisms giving rise to the effect are present in the early stages of visual processing. Low-level aftereffects such as figural aftereffects (Sutherland, 1954; Köhler and Wallach, 1944), the TAE, motion aftereffects (Addams, 1834), size aftereffects (Blakemore and Sutton, 1969), position aftereffect induced by motion (Whitney and Cavanagh, 2003), etc., are generally expected to arise in the early visual system, and specifically the primary visual cortex (V1), which contains neurons selective for these patterns.

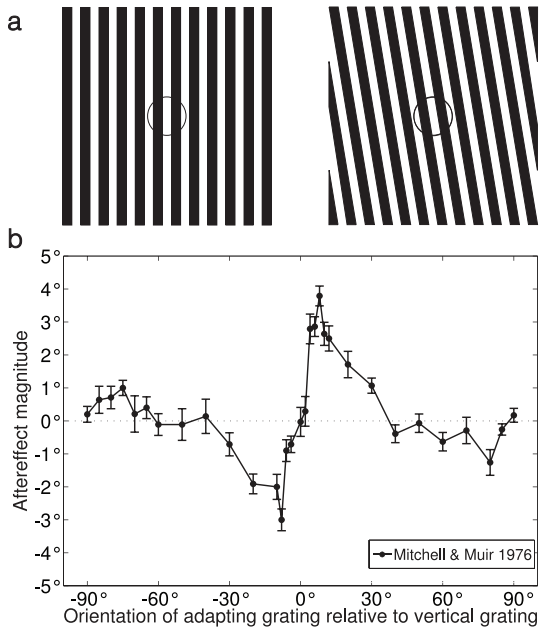


Figure 2.6: TAE experiment stimulus and results. **(a)**: Pattern of vertical grating (left) and tilted adapting grating (right). On the right pattern, fixate your gaze on the circle at the centre for at least thirty seconds, moving your eye slightly inside the circle to avoid developing strong afterimages. Then fixate on the left pattern. The vertical lines should appear to be tilted slightly clockwise. This is called the direct tilt aftereffect. On adaptation trials, the initial adaptation time is 180 seconds, and then 10 seconds in each trial; the presentation duration for the test stimulus is 2 seconds (Mitchell and Muir, 1976). Source: Bednar and Miikkulainen (2000) **(b)**: TAE curve measured in a human psychophysical experiment by Mitchell and Muir (1976). Adapting patterns similar to the test pattern lead the test pattern to be misperceived as more different than it is, while less similar adaptation patterns have little or even negative (attractive) effects.

2.4 Previous models of face and object processing

As data has become available about the physiology and psychophysics of face processing and face aftereffects, many theories and models have been proposed to account for how they are implemented neurally. The most relevant of these models, from the abstract framework of face space to more concrete connectionist models, are described in this section.

2.4.1 Face space theory

Face space theory is a general cognitive model for representing faces. Valentine (1991) first proposed the idea of a multidimensional space, where individual faces are encoded as a point. Individual faces can be discriminated in the space by their differences along each of the dimensions, such as facial distinctiveness, gender, race, etc.

This theory for face representation is simple and general, though highly abstract. It was proposed as a way to unify a series of models that were based on the assumption of a stored abstracted face prototype. Valentine (1991) suggested that the effects of facial distinctiveness, inversion, and race could all be explained by the stored information of the population of faces previously encountered. These ideas will be explored in the next section.

2.4.2 Norm-based and exemplar-based models

In the area of cognitive face coding, two types of theories — "norm-based" and "exemplar-based" — have existed in parallel for decades.

Early work by Valentine and Bruce (1986a,b) found that highly distinctive faces are recognised faster in a familiarity decision task but classified as faces more slowly than are typical faces. Rhodes et al. (1987) prepared veridical line drawings of faces, a norm face generated by averaging the position of the co-ordinates of 169 specified points across several faces, face caricatures generated by increasing the differences between

a face and a norm, and anti-caricatures generated by reducing the differences between a face and a norm. They found that the recognition speed of them was: face caricatures were faster than veridical line drawings, which were faster than anti-caricatures. But there were no differences in recognition accuracy. These studies led to a suggestion of a holistic encoding process in which the distinctive aspects of faces are encoded by comparison to a norm face, termed as a "norm-based" coding theory.

Rhodes and Jeffery (2006) proposed a two-component norm-based model to specifically account for face identity aftereffects. They postulated two pools of neural populations. The population tuning curves for these neurons have complementary values for a given face space dimension (shown in Figure 2.7). A face's value along that dimension is coded by the relative response of the paired populations; the norm face is coded by equal responses. When adaptation occurs, the activity of the pool that responses more strongly to the adaptation face is suppressed, and as a result, the equal point (face norm) moves towards the direction of the adaptation face. This process applies similarly for all the other dimensions of the face, with the result that viewing a face biases perception towards the opposite identity. Predictions from this model are consistent with the data from experiments in human face-selective brain areas (Loffler et al., 2005) and measurements of face-selective neurons in monkeys (Leopold et al., 2006).

Goldstein and Chance (1980) first proposed the alternative exemplar-based theory. They hypothesised that a face is recognised by using a "schema", and explained that with development, adults perform better (more accurate) than children in face recognition of other-race and inverted races due to better use of the schema, but become worse in efficiency (speed) due to increasing "schema rigidity". Diamond and Carey (1986) showed that the recognition of experts in another stimulus class (dog) is as adversely affected by inversion as is face recognition. They suggested that the exemplars should have a common configuration but subtle spatial differences, and the observers have sufficient expertise to distinguish them. As summarised in Valentine (1991) and Valentine

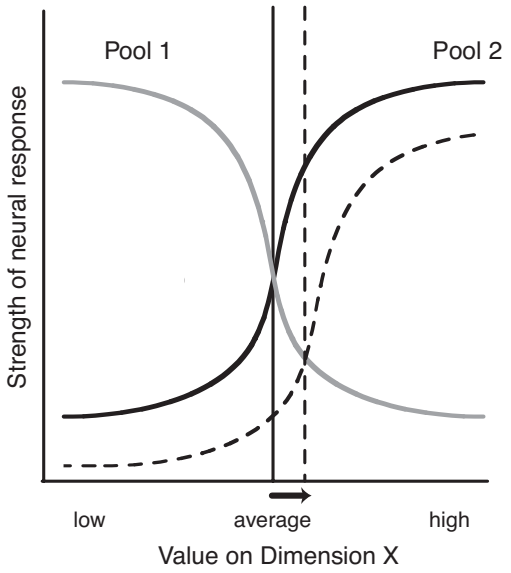


Figure 2.7: Two-component norm-based adaptive coding model for facial identity. Each dimension of face space is assigned to two pools of neurons. The two pools code complementary values on dimension X. The norm face is represented by equal response of the two populations. Exposure to an adapting face with a high value on dimension X will shift the perceived average towards the adapting face (dotted vertical line) and make the previously norm face take on the opposite identity. The aftereffect is measured as the reduction in response of the pool 2 neurons, which respond strongly to the adapting face. Source: Rhodes and Jeffery (2006).

and Endo (1992), exemplar-based theories usually have "multichannel" components, which represent the tuning curves of a wide variety of neurons. Each neuron is selective for a small range of faces in a dimension, and thus an individual face is coded by which neurons respond most strongly to it.

These two types of theories can both be implemented under the general framework of "face space"; they differ mainly in whether neuronal preferences along a face space dimension occur in pools or have widely varying values, and whether the neurons have localised or open ended responses. In recent years, many studies on face aftereffects have found evidence supporting the norm-based theory. Studies of human face gender, identity and distortion aftereffects (Webster et al., 2004; Little et al., 2005; Leopold et al., 2001; Jeffery et al., 2010) and physiological studies of the FAE in monkeys (Leopold et al., 2006) have led to an emerging consensus that faces are processed with norm-based encoding. However, predictions from exemplar-based models have not been tested extensively.

However, recent work by Chen et al. (2010) found face viewpoint aftereffects consistent with the exemplar-based theories. As illustrated in Figure 2.8a, they measured the angular tuning function of the face viewpoint aftereffect (Fang and He, 2005) and showed that with the adapting angle increasing from 0° to 90°, the aftereffect curve was not increasing monotonically, but increased and peaked at 20°, and then gradually decreased. This result was consistent with their proposed multichannel exemplar-based model as shown in Figure 2.8b. Their model also suggested tuning curves with very broad bandwidth. Moreover, Oruç and Barton (2010) devised a novel face adaptation method involving changes in contrast thresholds for face recognition, and found that recognition threshold had non-monotonic change at long adaptation durations. Their psychophysical result can be accounted for by a multichannel exemplar-based model. Although this work mainly focused on the dynamics of aftereffects, it still suggests an alternative theory for face adaptation in general. Even for the tested gender aftereffects described by Webster et al. (2004), one may argue the range of faces used in the

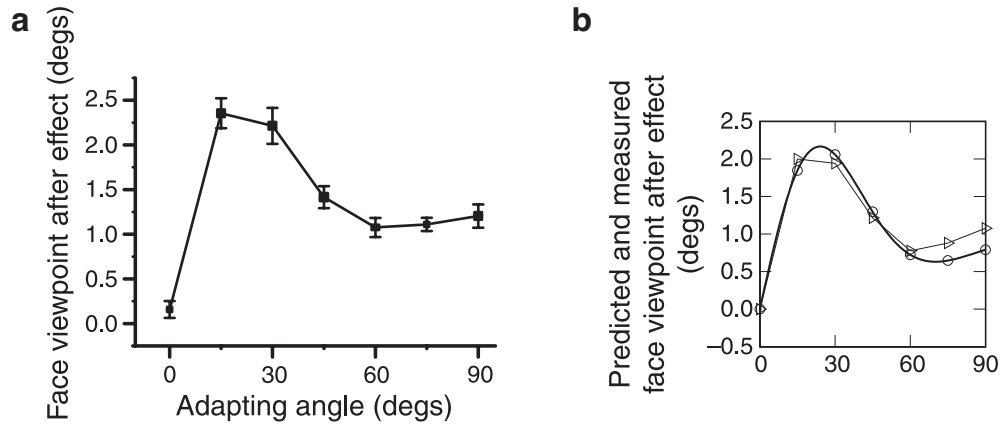


Figure 2.8: Experimental results and model for the face viewpoint aftereffect. **(a)**: Magnitude of the face viewpoint aftereffect plotted as a function of adapting angle. Positive values denote a repulsive aftereffect. Error bars denote 1 SEM calculated across subjects; Angular tuning function of the face viewpoint aftereffect predicted by a exemplar-based multichannel model (bold line) and measured in Experiment 1 (feint line). In the model, the adaptation effect manifests as a combination of response inhibition around and bandwidth broadening away from the adapting view. Source: Chen et al. (2010).

experiment was much smaller than in the real life (i.e., much more masculine or feminine faces may be encountered in daily life). Thus, given the broad tuning found so far, whether such aftereffects will be S-shaped like low-level effects, or monotonically increasing as predicted by the norm-based theories, is not known from the data so far. Thus despite this emerging consensus in favour of norm-based encoding, how faces are represented by neurons is far from being clear.

Note that in a trivial sense, a norm-based opponent model would presumably also lead to an S-shaped curve, with a stimulus sufficiently far from the norm leading to a reduced effect. At some point, a suitably extreme face (e.g. with eye separation much wider than the height of the face) would no longer be recognized as a face at all, so that neither of the pools of neurons would be activated, leading to no effect. However, for the norm-based model the reduction in response would presumably occur at or near the edge of representable stimuli, while for the multichannel model the reduction can occur well within the representable range (as in the model and in the data in this thesis).

More details on the abstract models implemented in this thesis will be described in Section 4.3, and a discussion comparing with existing work can be found in Section 7.1.2. While the abstract models can explain the face aftereffects in a general framework, they cannot be related to detailed information about neural activities about how the aftereffects arise based on specific visual stimuli. Connectionist models with a more physiological basis will be introduced in the next section.

2.4.3 Connectionist models

As neurophysiological researchers have found more evidence of how visual neurons are organised and code for objects or faces, many theoretical models have been proposed to show how neurons can be connected to achieve plausible recognition.

A series of papers by the Rolls group (Rolls and Milward, 2000; Stringer et al., 2006) describe the VisNet model of the ventral visual pathway, implementing basic invariant face/object recognition. VisNet is a hierarchical feedforward network to investigate invariant visual object recognition in high-level visual systems. The hierarchical layers represent cortical regions in the ventral pathway, and the final region outputs a view-independent, translation-independent, or size-independent object representation. The receptive field (RF) size increases along the pathway, converging to the final layer, whose neurons have an RF covering the entire input layer. The neurons in each layer are competitive through mutual inhibition, and the neuron's activities are updated by a trace learning rule (Földiák, 1991) to learn from the changing inputs. Overall, although the representations obtained are only partially invariant, the multiple-layered convergent RF architecture is plausible, and demonstrates the idea that with gradually expanding RFs, neurons in upper layers can achieve more abstract processing.

Riesenhuber and Poggio (1999) proposed a hierarchical model HMAX for invariant recognition, where objects are represented with respect to their original viewing conditions. Their model proposed that the similarity between an input and memory of previous inputs is computed using a collection of radial basis functions, each centered

on a meaningful feature in the image. The model is arranged in alternating simple and complex cell layers, preferring either linear summation (simple cells) or taking the maximum of their set of inputs in a preceding layer (complex cells). Through these operations, pattern specificity and invariance to translation are provided by pooling over afferents tuned to different positions, scales, etc. Later models inheriting the idea of HMAX added a sparse coding scheme (Reddy and Kanwisher, 2006) to achieve invariance. Complex stimuli such as face identities can then be sparsely represented by a small number of neurons that are highly selective for them.

Other types of models focused on more specialised aspects of face processing and adaptation.

For example, Dailey and Cottrell (1999) studied the neural mechanism of face processing specialisation and clustered organisation. They proposed a model composed of two "experts", implemented as linear classifiers, between which competition is mediated by a gating network. One of these receives higher spatial frequency information, the other lower spatial frequency information, but neither contained any specific hardwiring or feature selectivity for faces. Even with only this very weak bias, after repeated presentation of face and non-face patterns, one module reliably became face specialized and the other non-face specialised. They concluded that there is no need for an innately specified face-processing module, and that face recognition is only special because faces form a remarkably homogeneous category of stimuli.

The Sohal and Hasselmo (2000) model addressed neural adaptation caused by familiar stimuli. Such neural adaptation might be the basis of high-level adaptation that leads to face aftereffects. It used an IT representation to model neural familiarity effects (declining IT neuron responses with familiarity to visual stimuli). This model features a self-organised competitive input region and an IT region, plus a module of cholinergic neurons that modulates the activities in the input and IT regions. In this model, a cholinergic neuron is most active in response to novel stimuli in the input, and suppressed in response to a familiar stimulus. Through such modulation, non-

overlapping representation and declined activity of neural responses are obtained in IT region. Here the idea of cholinergic modulation may be applied to the feedback network to produce neuronal adaptation, which would be the basis of face aftereffects.

Despite these interesting but abstract models, there are few models directly addressing the neural mechanisms of face adaptation implemented on a stimulus-driven neural network. The next section describes such a model previously implemented for early visual cortical areas, but general enough to be applied across the cortical hierarchy and tested for face aftereffects.

2.4.4 LISSOM model

This thesis aims at investigating whether face adaptation can be achieved through a general neural mechanism performing both low-level and high-level processing. Consistent with the findings of Leopold et al. (2006), the models in this thesis will be built using a neural network implementing visual cortex levels from V1 to IT, linked to psychophysical results. The same low-level mechanisms used in the V1 level will be applied in IT as well. This laterally interconnected synergetically self-organising map (LISSOM; Miikkulainen et al., 2005) model will be introduced in this section and used throughout the thesis. LISSOM was chosen because it is relatable to the underlying neural architecture of the cortex, and it has been shown previously how this structure will lead to realistic low-level aftereffects (Bednar and Miikkulainen, 2000). The mechanisms in LISSOM are also very general, and thus it is straightforward to extend it to model face perception.

The LISSOM network primarily shows how topographical maps can develop in V1. It is a hierarchical neuronal network architecture for the simulation of cortical regions. Each region is represented by a competitive network with local excitation, where Hebbian learning leads neurons to become similar to their neighbours yet different from more different neurons, gradually forming a self-organising map. A diagram of the simplest version of the model is shown in Figure 2.9.

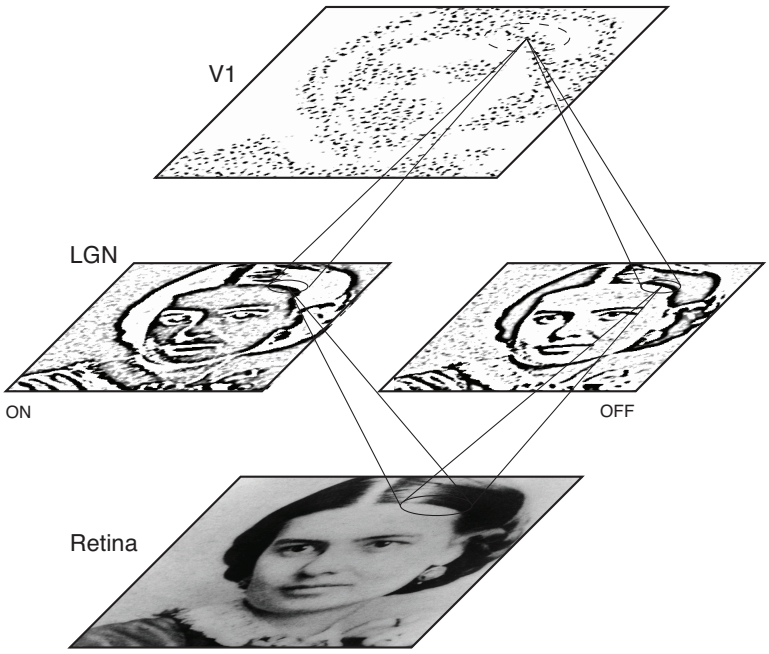


Figure 2.9: Basic LISSOM model of the primary visual cortex. V1 is represented by a two-dimensional array of neurons. These neurons receive input from the retinal receptors through the On/Off channels of the LGN, and from other columns in V1 through lateral connections. The solid circles and lines indicate the receptive fields of neurons in the LGN and V1. The dashed circle in V1 indicates the lateral connections of a V1 neuron. The LGN and V1 activation in response to a sample input on the retina is displayed in gray-scale coding from white to black (low to high). The V1 responses are patchy because each neuron is selective for a particular orientation, and only particular orientations exist at each location in this image. Source: Miikkulainen et al. (2005).

In this model, neurons in two LGN sheets have opposite shape of difference-of-Gaussian receptive fields. The activity of an LGN neuron is defined by

$$\xi_b = \sigma\left(\gamma_{\mathbf{L}} \sum_a \chi_a L_{a,b}\right), \tag{2.1}$$

where χ_a is the activation of cell a in the receptive field of b, $L_{a,b}$ is the afferent weight from a to b, and $\gamma_{\mathbf{L}}$ is a constant scaling factor. The squashing function $\sigma(\cdot)$ is a piece-wise linear activation function, zero below a threshold, and linear until a saturation cut off. The activation of neuron i in the V1 at time t is defined by

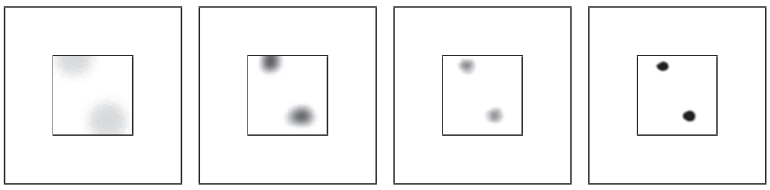
$$\eta_i(t) = \sigma\left(\sum_\rho \gamma_\rho \sum_{j \in \mathbf{F}_\rho} X_j(t-1) w_{ji}\right), \tag{2.2}$$

where the index ρ indicates a type of incoming connection field ($\mathbf{F}$; afferent, lateral excitatory or lateral inhibitory), $X_j(t-1)$ is the activation of unit j in that connection field, and w_{ji} is the weight from that unit to unit i. The sign of the scaling factor γ_ρ is positive for afferent and lateral excitatory connections, and negative for lateral inhibitory connections. In the cortical sheet (beyond retina and LGN), a neuron not only receives afferent inputs from preceding sheets, but also long-range inhibitory and short-range excitatory lateral connections from neighbouring neurons on the same sheet. These inputs are processed recurrently until the activity stabilises or "settles" into discrete blobs of activity. An example of the initial and settled V1 neuron response in the model is shown in Figure 2.10. Activated neurons then learn to represent the input that was present when they were activated, leading to maps organised to represent a feature dimension (such as orientation).

A divisively normalised Hebbian learning rule is used to update the above weights:

$$w'_{ji} = \frac{w_{ji} + \alpha X_j \eta_i}{\sum_k (w_{ki} + \alpha X_k \eta_i)}, \tag{2.3}$$

where w_{ji} is the current afferent or lateral connection weight from j to i, w'_{ji} is the new weight to be used after the end of the settling process, α is the learning rate for this type



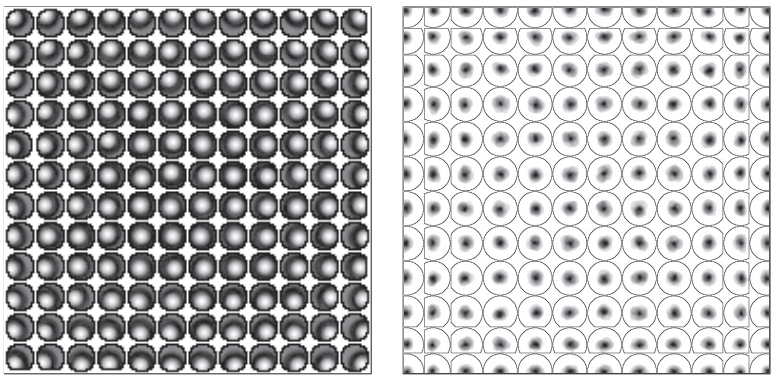
(*a*) Iteration 0: Initial V1 response | (*b*) Iteration 0: Settled V1 response | (*c*) 10,000: Initial V1 response | (*d*) 10,000: Settled V1 response

Figure 2.10: Example V1 responses. At iteration 0, the V1 map initially responds broadly and unspecifically to the input patterns (a); The lateral connections focus the response into discrete activity "bubbles" in (b), and connections are then modified. After 10,000 input presentations and learning steps, the initial and settled V1 responses are more focused, forming a sparse representation of the input (c and d). Source: Miikkulainen et al. (2005)

of connection, X_j is the presynaptic activity after the current settling step, and η_i is the activity of neuron i after settling. Afferent inputs, lateral excitatory inputs, and lateral inhibitory inputs are normalised separately, to ensure that the relative contribution from each type remains balanced. The normalisation process redistributes the weights so that the sum of each weight type for each neuron remains constant. The learning rates α are reduced gradually during the course of simulation in order to develop a smooth organisation and well-tuned receptive fields. Learning will continue indefinitely, but the weights change very little after 10-20,000 iterations, at which point the model is considered fully developed.

In a LISSOM simulation driven by input stimuli of orientated Gaussian bars at random position, the LISSOM network can eventually self-organised to a map representing the retinotopy — the only feature the input space has. An example of the self-organised afferent and lateral V1 weights for this retinotopy map is shown in Figure 2.11.

Note that in the LISSOM model, while the weights of lateral connections follow the Hebbian learning rule during model simulation, the maximum spatial extent of the lateral excitatory connections is gradually reduced, following a preset schedule (Bed-



(*a*) Afferent (ON - OFF) (*b*) Lateral inhibitory

Figure 2.11: Self-organised afferent and lateral weights across V1. In this V1 array, each circle represents a neuron. The afferent map (a) has learned to represent the retinotopy feature space, with the weight patterns systematically moving outward away from the centre of the receptive field. The weights are expanding the entire area. The lateral connections in (b) approximate the input space (difference of Gaussians). Source: Miikkulainen et al. (2005)

nar and Miikkulainen, 2000). This shrinking helps ensure that the map becomes well organized over time, by establishing global and then local order, but is not necessary in later versions of the model such as GCAL (Antolik, 2011).

Even though it is primarily a developmental model, LISSOM has also shown how adaptation can be implemented by lateral inhibitory learning alone, and how the TAE can arise from such neural adaptation (Figure 2.12). The architecture of the LISSOM TAE model is the same as described before, featuring sheets of photoreceptors, RGC/LGN neurons, and V1. Inputs to the model are oriented Gaussian bars. An example of vertical Gaussian bar input is shown in Figure 2.12a. After the model was trained for 20,000 iterations, an orientation preference map was measured, shown in Figure 2.12b, with colours standing for orientations as shown in the colour chart. Each point in this map represents the preferred orientation of the neuron in that cortical sheet position. Figures 2.12c and 2.12d show how to estimate the orientation of an input by using this map. The perceived orientation of a new input is predicted by the vector sum of neurons' preferred orientations. As illustrated, this orientation was perceived as -1.3^{deg} by the settled activities in Figure 2.12d, very close to the veridical verti-

cal orientation. Adaptation was then simulated by presenting a fixed vertical Gaussian bar for 90 iterations of adaptation. The perceived orientations were computed before and after adaptation, and the aftereffect was estimated as the difference between pre-adaptation orientations and post-adaptation perceived orientations. The predicted aftereffects are shown as a solid line in Figure 2.12e, against the human experimental data by Mitchell and Muir (1976) as dashed line. It can be seen that the modelled result is very close to the human experimental data. Thus, the LISSOM network provides a solid and reliable basis for simulating visual aftereffects at a neural level, in the context of a general explantion for neural development. Because the model is very general, it can be modified easily to study high-level adaptation, as will be done in this thesis.

One of the most important features of LISSOM and related models is its underlying simplicity. Every neuron in a cortical sheet is purely driven by a self-organising process, and updates its activities by specific learning rules. The behaviour of a single unit is simple, but the collective activities of neurons on a cortical sheet or cross-cortical sheets become complicated and highly organised, leading to a well-organised V1 orientation preference map as shown in Miikkulainen et al. (2005), as well as realistic aftereffects. Overall, LISSOM aims to simulate the brain without any circuits or mechanisms inherently specialized for particular stimuli or sensory modalities; instead the cortical activities emerge solely by each neuron's simple mechanism and the connections between them, with parameters that are set based on sensory experience.

Overall, LISSOM is based on the idea that processing in visual cortical areas is not genetically determined and hardwired, but primarily the result of external stimuli affecting neural plasticity mechanisms. Most LISSOM models include only the photoreceptors, RGC/LGN and V1, but some have focused on face perception in human newborns (Miikkulainen et al., 2005). Low-level studies show behaviour consistent with optical imaging and physiological experiments on orientation preference and selectivity, as well as showing how tilt aftereffects can arise from short-term intra-cortical synaptic adaptation.

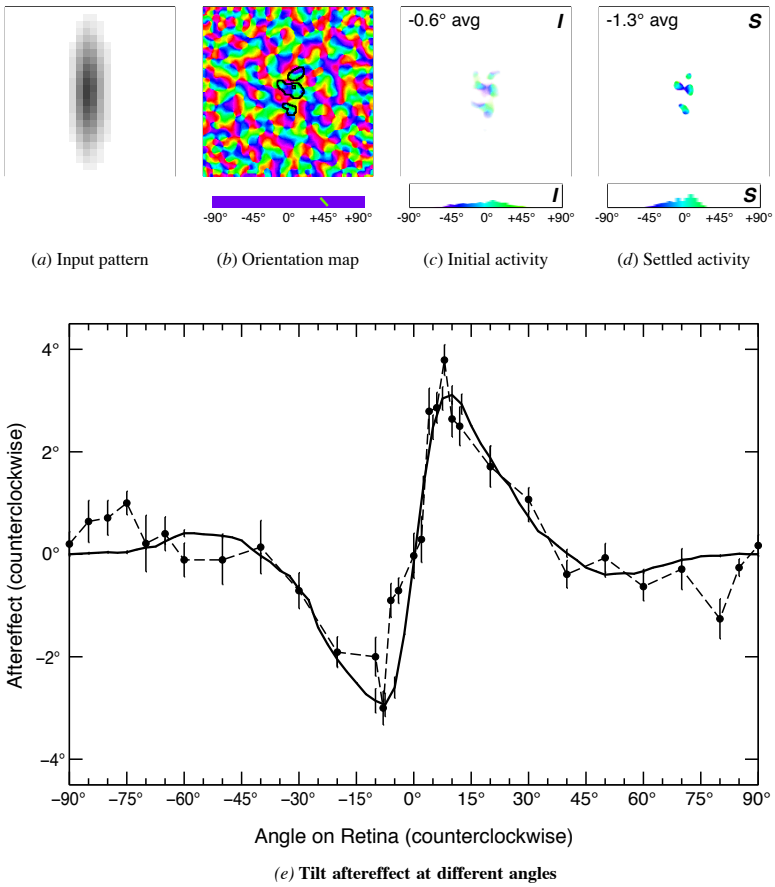


Figure 2.12: Simulation and result on the TAE modelling. In a trained LISSOM V1 network, the orientation preference map (b) is used to measure the orientation of the input stimulus (a). Plots (c) and (d) show the initial and settled response of V1 neurons, with the colours showing the preferred orientations of the neurons most responsive to the stimulus. It can be seen from (d) that the vector sum of the preferences approximate a vertical orientation, very close to the orientation of the input stimulus. The measured orientation of vertical angles under all adaptation conditions are calculated and plotted in (e), as aftereffect values (post-adaptation orientation minus pre-adaptation orientation). The resulting aftereffect curve approximates the experimental data by Mitchell and Muir (1976). Source: Bednar and Miikkulainen (2000)

To summarise the main conclusions of this chapter, the neural mechanisms of face processing are largely unknown, and techniques for investigating these mechanisms in human are very limited. Instead, psychophysical researchers have examined a series of face aftereffects, which give clues about the underlying mechanisms of face processing. Many theories have been proposed to account for these effects, but most of them are formulated at an abstract level that is difficult to relate to neural processing. Very few concrete neural models have been used for studying face aftereffects. The previously introduced LISSOM model is a good starting point, as it has is developmental model with a general architecture, simulation of individual neurons, and has previously demonstrated realistic low-level aftereffects. In this thesis, the limitations of existing work will be overcome by building an extended LISSOM model for face processing (Chapter 3, 4 and 6), measuring face gender aftereffects (Chapter 5), and comparing the experimental and modelling results.

Chapter 3

Modelling face identity aftereffects

This chapter first briefly reviews the previous experimental work on face identity aftereffects and states the motivation for modelling it. Then the details of the modelling work are introduced, including the model architecture, face generator, decoding methods and the aftereffects simulation. The methods for both duplicating the existing face aftereffect experiment and for illustrating the full picture of the multi-dimensional face perception shift are described. The result of the model output is shown and compared with the experimental work. The results suggest that existing cortical models can replicate the experimental results for face aftereffects, and suggest simpler models and experiments that can be used to test prediction of these models.

3.1 Introduction

Physiological and imaging studies have so far provided limited evidence about how face processing might actually be achieved in the human ventral visual stream. Yet psychophysical studies provide interesting clues that could lead to a more detailed explanation of this ability. As discussed in section 2.2, Leopold et al. (2001) found a face identity aftereffect — brief exposure to an individual face (i.e., adaptation face) generates significant and systematic misjudgements in the subsequent perception of face identity. They showed that participants' face recognition performance was facilitated

if the adaptation and test face are on the same identity dimension, but impaired if on different identity dimensions.

This striking aftereffect provides a crucial link between the computational modelling of face processing and psychophysical experiments, and serves as a strong constraint to the visual system models. As discussed in section 2.4, many computational models have been proposed to explain how neuronal circuitry could support human-like face and object recognition. It is therefore crucial for a plausible theory to have a testable case of how the high-level behaviours are simulated, based on a hypothesised low-level neural mechanism. Face identity aftereffects are such a suitable case, and thus it is worthwhile modelling them based on existing neural models. This is the motivation for this chapter's modelling work, which can be a valuable constraint to help understand how the visual system works on a neural level.

Many psychological studies have considered a face space framework and proposed a norm-based theory to account for how face space shifts neurally during adaptation (see section 2.4.2 for details). While the norm-based theory can well explain face aftereffects, it requires mechanisms seemingly very different from those that appear to be implemented in well-studied low-level visual areas, e.g., for relating responses to a specific "mean face". It will be more intuitive if such a high-level face space representation can arise from the same basic low-level neural operations found in e.g., V1. Many models address this idea (for a review, see section 2.4). In particular, this chapter proposes a conceptual model based on LISSOM. Described in detail in section 2.4.4, LISSOM models a self-organising process of neural development for afferently and laterally connected hierarchical neuron sheets, with Hebbian learning of connections. LISSOM has shown matching physiological results to optical imaging of topographical structure and single-unit responses in V1. Based on this well-established modelling architecture, it will be helpful to see the topographical structure of neurons in a higher visual area, how this structure could give rise to the psychophysically measured face space, and how this structure adapts during face aftereffects.

In all, this chapter aims to duplicate the main results in the human experiment described by Leopold et al. (2001) and tries to reveal more information about the neural mechanisms of face adaptation.

3.2 Methods

In this section, the details of the modelling face identity aftereffects will be described, including how the model was organised, different types of face generators, two kinds of decoding methods, and the process of aftereffect simulation.

3.2.1 Model architecture

It is widely accepted that the ventral visual stream plays a key role in object recognition (Ungerleider and Haxby, 1994). The task of object recognition is thought to be performed through this hierarchical cortical organisation, starting from the retina and simple cells in V1. Higher visual areas such as the inferotemporal cortex (IT) in this pathway have broadly tuned cells that are robustly invariant to complex object transformations, such as scale and position changes (Bruce et al., 1981). Yet, how this process works exactly to achieve the remarkable tasks of object recognition is largely unclear physiologically, psychologically, and theoretically (for a review, see Peissig and Tarr, 2007).

Because the aim of this chapter is not to achieve invariant recognition, the models considered in this chapter bypass the hierarchical cortical organisation implemented in models like HMAX and VisNet (Section 2.4.3). Instead, the models use a single cortical sheet named the "face-selective area" (FSA) consisting of face-selective neurons with large receptive fields, in order to match the underlying architecture of the ventral stream. This simplification bypasses all invariance processing and assumes that inputs to the FSA have been properly processed to a normalised size, orientation, position, etc. This treatment makes it easier to focus the work on face aftereffects. To meet this

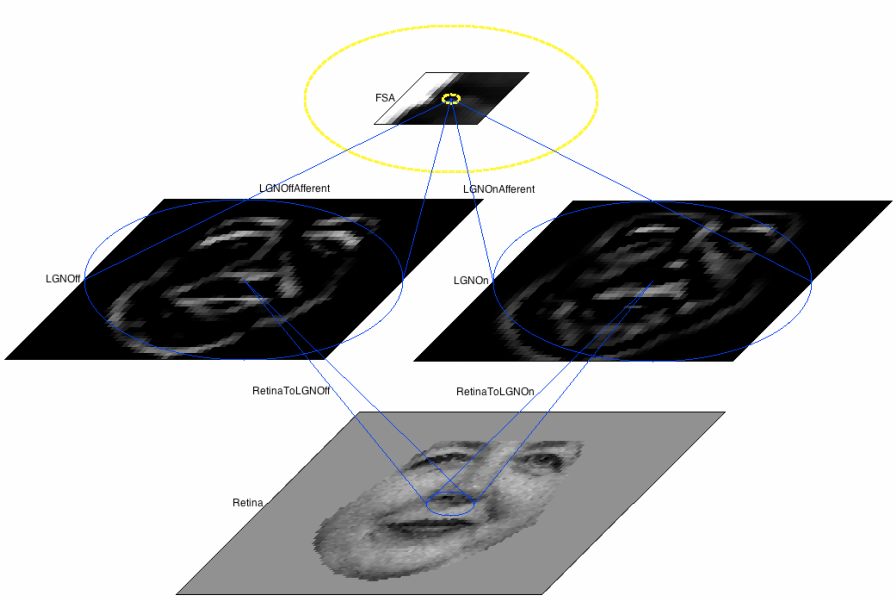


Figure 3.1: Diagram of simple FSA model in LISSOM. **Top**: face-selective area (FSA) sheet as representation of face space; **middle**: LGNOn and LGNOff sheet as a simulation of the pathway from the photoreceptors to the cortex, including the retinal ganglion cells and the lateral geniculate nucleus of the thalamus; **bottom**: photoreceptors input sheet.

requirement, all the input faces in this chapter were registered with key points, and they all had a normalised size and clear background. The original faces as shown in Figure 3.2 and their registration (at the centre of two eyes) were both produced for the dataset by Nordstrøm et al. (2004). For more details about face registration and normalisation, see the next section 3.2.2. Models with hierarchical arrangements of cortical sheets as in the ventral stream will be explained in Chapter 6. A diagram of the initial model is shown in Figure 3.1. The key theory of this model has been explained in Section 2.4.4.

In other LISSOM models, a V1 sheet takes the place of the FSA in this model. Even though both FSA and V1 occupy similar positions in the network architecture (i.e., both directly receiving inputs from the LGN sheet), the configuration of the FSA sheet is different from V1. The receptive field of a neuron in FSA covers all LGN cells, and thus every neuron in the FSA can learn the entire stimulus. This setting corresponds to the large receptive field size shown in many studies and used in many

existing models. The learning rates of afferent and lateral connections were also all higher than those in the standard LISSOM V1 model, as a method of speeding up the simulation. The learning parameters were further tuned so that after 20000 iterations, a distinctive representation arises in the FSA for each face stimulus.

The specification of FSA model in this chapter is similar with the FSA model described in Table C.1 in Miikkulainen et al. (2005), except for the radius of afferent connections (r_A), the initial radius of lateral inhibitory connections (r_{E_i}), and the learning rates of all afferent and lateral connections (α_{A_i}, α_{E_i} and α_I). Specifically, r_A is 1.2 times of the value in Table C.1 (Miikkulainen et al., 2005), r_{E_i} two times of the value in Table C.1 (Miikkulainen et al., 2005), and α_{A_i}, α_{E_i} and α_I all 1.2 times of the value in Table C.1 (Miikkulainen et al., 2005).

3.2.2 Face generator

Much literature on face psychophysics has suggested that a "face space" underlies the processing of faces. Although there is no fixed specification of such a space, it should map the range of typical input stimuli, and be generated by systematic multi-dimensional parameters, each characterised by some feature in which real faces vary.

In order to see how the model behaves in the simplest case, a one-dimensional face generator was first implemented and tested. Leopold et al. (2001) first demonstrated face adaptation aftereffects elicited by their face stimuli, which were generated by an active appearance model (AAM) building three-dimensional faces (Blanz and Vetter, 1999). These faces are linear combinations of shape and texture eigenface components. Such an eigenface basis may correspond to dimensions in the face space observed in humans (Meytlis and Sirovich, 2007). However, despite of several requests, this tool has not been made accessible to researchers outside their group. As a workaround and simplification, a 2D implementation of AAM (Cootes et al., 2001) was used. This method uses a set of face images as a training set, and builds a model of facial appearance. Then two face identities are synthesised using this set. Between these two

modelled identities, a face continuum is generated characterised by a parameter. The faces in this continuum form a one-dimensional face space. Details of this procedure are described below.

The face dataset from Nordstrøm et al. (2004) (data can be downloaded from `http://www2.imm.dtu.dk/aam/datasets/datasets.html`) was used to construct the appearance model. This dataset is composed of 40 still images of different face identities. Nordstrøm et al. manually annotated each face using 58 landmarks at eyebrows, eyes, nose, mouth, and jaw. These key positions outlined the main facial features. A common coordinate system was set up so that each face can be represented by these key position coordinates, together denoted as vector $\mathbf{x}$. A principal component analysis (PCA) was applied to these face position vectors and a face shape $\mathbf{x}$ can be approximated by

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s, \tag{3.1}$$

(Cootes et al., 2001) where $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}_s$ is a set of orthogonal modes of shape variation and $\mathbf{b}_s$ is a set of shape parameters. Likewise, each face texture needs to be approximated from the mean texture. To achieve this, the key points of each face image were warped to match the mean shape using Delaunay triangulation, and then a set of shape-normalised face texture patches was obtained. Mean face texture $\bar{\mathbf{g}}$ was estimated by applying a recursive minimisation procedure on this set of face texture patches. PCA was then applied to these normalized face textures and a face texture $\mathbf{g}$ was approximated by

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{3.2}$$

(Cootes et al., 2001), where $\mathbf{P}_g$ is a set of orthogonal modes of texture variation and $\mathbf{b}_g$ is a set of texture parameters. Thus, the shape and texture of a face can be characterised by vectors $\mathbf{b}_s$ and $\mathbf{b}_g$. There are correlations between shape and texture information, and it would be more helpful if a face could be described by a combined appearance

parameter vector. By applying a PCA on $\mathbf{b}_s$ and $\mathbf{b}_g$, a combined model can be obtained

$$\begin{pmatrix} \mathbf{W}_s\mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_s \\ \mathbf{Q}_g \end{pmatrix} \mathbf{c}, \tag{3.3}$$

(Cootes et al., 2001) where $\mathbf{W}_s$ is a diagonal matrix of weights for each shape parameter, $\mathbf{Q}_s$ and $\mathbf{Q}_g$ are eigenvectors, and $\mathbf{c}$ is a vector of combined appearance parameters. This way, the shape $\mathbf{x}$ and texture $\mathbf{g}$ of a face can be approximated by

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s\mathbf{W}_s\mathbf{Q}_s\mathbf{c}, \mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g\mathbf{Q}_g\mathbf{c}. \tag{3.4}$$

(Cootes et al., 2001).

By using the transformation described in equation (3.3) and (3.4), the shape and texture parameters $\mathbf{b}_s$ and $\mathbf{b}_g$ are now replaced by a single combined appearance parameter vector $\mathbf{c}$. This way, a new face can be synthesised by generating texture $\mathbf{g}$ and warping it using the key points $\mathbf{x}$. In order to construct a one-dimensional face space, two identities (characterised by combined appearance parameter vectors $\mathbf{c}_0$ and $\mathbf{c}_1$) were synthesised. These two vectors should satisfy

$$\mathbf{c}_1 = f \cdot \mathbf{c}_0, \tag{3.5}$$

(Cootes et al., 2001) where f is the one-dimensional face space parameter - caricaturisation (i.e., morphing points between or beyond two faces). A continuum of faces can be generated by altering f values.

The overal lightness of the generated faces needs to be normalised before they are used as inputs to the model, in order to avoid dominance of particular lightness values during model training. In this chapter, each grayscale face image was L1-normalised. I.e., each pixel in a face image was divided by the sum of the values of all pixels.

Using the method described above, two visually distinctive faces were generated to train the LISSOM network as they provide the most information and might be the most likely to elicit distinctive neural responses. The LISSOM network was trained

Figure 3.2: Sample faces of one-dimensional face space and their activity patterns in FSA after training for 20,000 iterations. Face caricaturisation 0.0 and 1.0 were first synthesised, and then other faces were generated between or beyond them. **Row 1-2**: generated face caricaturisation -0.4 to 0.4 and corresponding FSA activity after training. **Row 3-4**: generated face caricaturisation 0.6 to 1.4 and corresponding FSA activity after training.

for 20,000 iterations with 10 caricaturisations (from -0.4 to 1.4 with 0.2 steps), 0.0 standing for one original face and 1.0 for another. Then the network was tested on a smaller caricaturisation interval (from -0.4 to 1.4 with 0.05 steps). Figure 3.2 shows the samples of the 10 faces in this continuum, and their activity patterns in FSA after training. Face caricaturisation 0.0 and 1.0 were first synthesised, and then other faces were generated between or beyond them.

This simple two-face-morph generator only showed one possible one-dimensional face space. The two faces were chosen arbitrarily, and thus it is hard to accurately judge results based on these. A more systematic approach is required to generate faces like those used in the study conducted by Leopold et al. (2001).

To achieve this, a more comprehensive face set (Hancock, 2000) was used to syn-

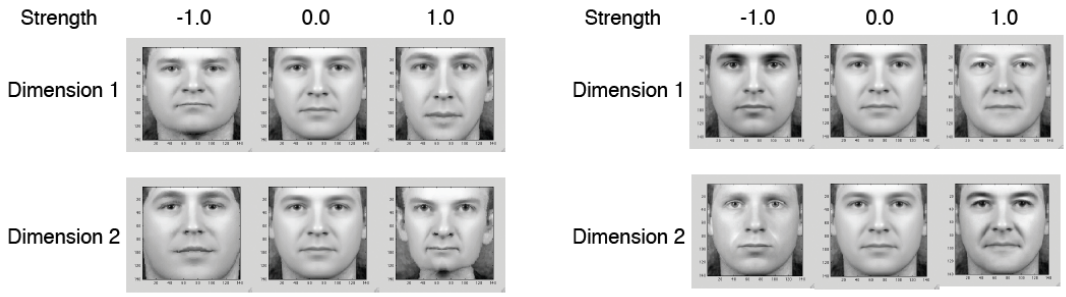


Figure 3.3: **Left**: shape dimensions of different strengths for the Hancock (2000) face generator. **Right**: texture dimensions of different strengths.

thesise an arbitrary face in terms of its texture and shape variations. This made it possible to duplicate Leopold et al. (2001). The general algorithm of this generator is very similar with the method described above, except for its large number of faces and richly annotated landmark points. The photographic images were produced by combining texture components ("eigenfaces") and shape components ("eigenshapes"). 206 faces were analysed to extract texture components. These faces were also annotated with over 200 controlling points and their x and y coordinates were analysed to obtain shape components. Thus, an original face can be restored by a linear combination of all the texture and shape components in appropriate proportions; likewise, a new face can be produced with a linear combination of certain proportions. These 206 original faces are all male and of the same race in order to avoid gender or race effects. In this thesis, these components are called "dimensions", not to be confused with the physical dimensions of space as in 2D and 3D renderings. Some examples of texture-only and shape-only dimensions are shown in Figure 3.3.

Theoretically, a face space has an unlimited number of dimensions, but probably only the first 40 or more dimensions have a meaningful impact on the face appearance (Meytlis and Sirovich, 2007). For practical reasons, the number of dimensions varied in this thesis was kept even lower — two and three. On the one hand, this is computationally tractable, and on the other hand visualisation of face space is straightforward with two or three dimensions. It should also be noted that this face generator is not as

sophisticated as the one used in the Leopold et al. (2001) experiment. Their approach incorporated a segmented morphable model that divides faces into independent subregions that can be morphed independently. This treatment made the dimensions in the face space more independent and makes the synthesised face more realistic, but as mentioned previously their generator has not been released to other researchers.

A three-dimensional face generator was used to generate LISSOM training inputs to simulate the Leopold et al. (2001) experiment. Two texture dimensions and one shape dimension were used to construct the face space, and hence the generated faces were linear combinations of these three dimensions (components). The proportional values for these three dimensions were from -6.0 to 6.0, denoted as (d_0, d_1, d_2). Face $(0.0, 0.0, 0.0)$ indicated the mean face. In the Leopold et al. (2001) experiment, a target face was used for human participants to discriminate among morphs of four faces and decide which one is most like the target face. Each of these four faces formed a trajectory in the face space, and morphing faces in a trajectory was denoted by a "strength" value. Accordingly, in the LISSOM simulation, a trajectory was defined by one of its end faces, e.g. $(-6.0, -6.0, 6.0)$, and the strength of a trajectory was defined as proportional to the three dimension values. For example, strength 1.0 face in this trajectory meant face $(-6.0, -6.0, 6.0)$, strength 0.5 meant face $(-3.0, -3.0, 3.0)$ and strength -0.6 meant face $(3.6, 3.6, -3.6)$. Another trajectory, for example, could be defined by face $(6.0, 0.0, -6.0)$. In this model, there were only three dimensions (in the Leopold et al., 2001 face generator there are many more), so the four trajectories were manually selected in order to make them as visually distinct to each other as possible.

After the model is trained, face preference maps can be computed for each dimension. Such a map shows each FSA neuron's preferred stimulus value (ranging from -1.0 to 1.0) in a dimension. Either in texture or in shape dimensions, close values mean neighbouring faces. These look alike in terms of the texture or the shape features. Therefore, if the model is well-organised by the input stimuli, the self-

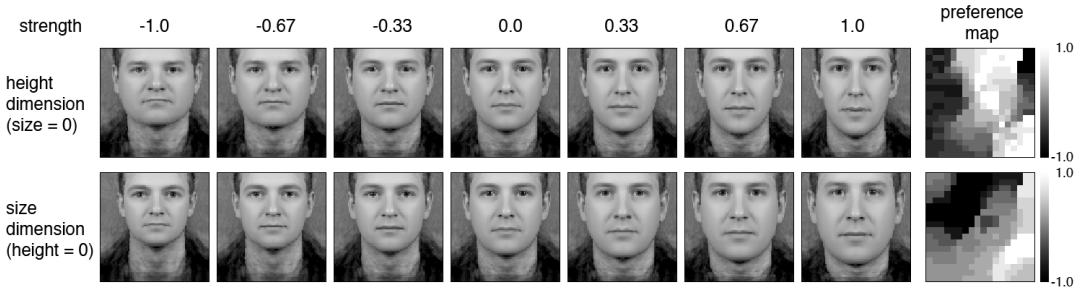


Figure 3.4: Face and preference map samples. **Top row**: sample faces and preference map of face height dimension (size = 0.0); **Bottom row**: sample faces and preference map of face size dimension (height = 0.0). Both vales are well represented in the map, and each neuron prefers a specific value along each dimension.

organising process should have clustered similar-looking faces to the neighbouring neurons, and thus the model should present a smooth (in grey-scale colour) preference map and gradually-changing receptive fields following certain directions in the FSA sheet. However, only the shape dimensions are well represented in the maps, in part because the basic LISSOM architecture inherently discards information relevant for texture processing (such as absolute brightness values, which are removed by the LGN processing). Therefore, in order to study the aftereffects in this model, the texture dimensions were dropped and only the shape dimensions were used for future study.

As a result, two of the most visually distinctive dimensions were finally chosen — face height and face size — to train the model and test face identity aftereffects. Figure 3.4 shows the sample faces of these dimensions and their preference maps after 10,000 iterations of training. It can be seen in Figure 3.4 that the preference maps for these two dimensions are smooth and represent a wide range of values along these dimensions allowing the model to be tested for aftereffects for these feature values.

3.2.3 Decoding method for duplicating the Leopold et al. (2001) experiment

With a LISSOM FSA model and one- and two-dimensional face generators at hand, it is now possible to decode the input information (single faces of different heights and sizes) from the model output. In this section, the following decoding methods will be described: population vector, maximum-likelihood (ML), support vector regression (SVR) and correlation comparison.The rationale for trying them will be discussed, and only correlation comparison method will be used for simulation.

Population vector decoding is a simple yet reasonable way to "read out" neuron responses. Bednar and Miikkulainen (2000) showed that population vector decoding in LISSOM V1 can read out stimuli orientation, and reliably demonstrate tilt aftereffects comparable with experimental results. They did this by averaging V1 responses weighted with each neuron's orientation preference. However, in that model, the population tuning curve for each stimulus was bell-shaped. In the model presented in this chapter, the shape of the population curve is less like a bell if the stimulus is more distant from the central caricaturisation (0.5), because of the effect of the limited non-cyclic input dimension (see Figure 3.5). As a result, for the stimuli around the extremes, the weighted average method is unlikely to compute the tuning curve peak as the perception.

It is also possible to compute a linear mapping between an FSA activity pattern and its corresponding stimulus. This can be an optimal linear estimator by which the stimulus can be accurately measured, given a trained face input. However, the resulting mapping is very sparse and limited to generalise over unseen inputs. When it is used as a linear transformation for unseen FSA responses, the results will be unpredictable. Unseen stimuli, and hence unseen responses, are crucial to adaptation aftereffects.

Instead of either of these two previous methods, two approaches were proposed here to overcome these issues. The first one is the maximum likelihood (ML) approach

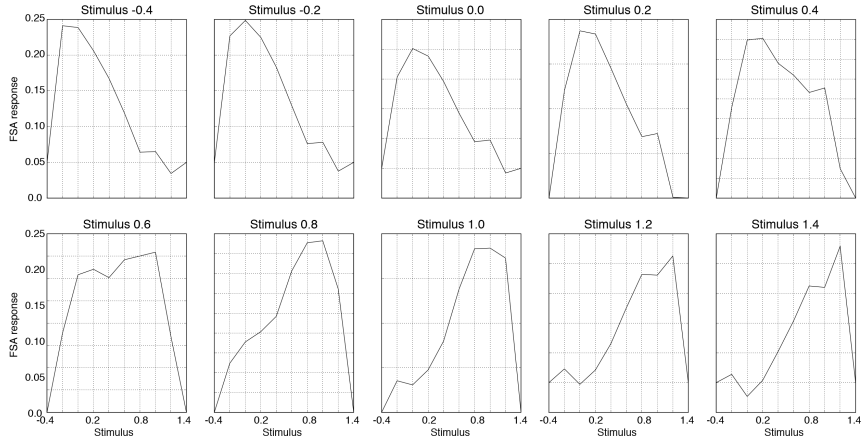


Figure 3.5: Population tuning curve for stimuli (caricaturisations -0.4 to 1.4). Each plot shows the histogram of FSA activities against each caricaturisation preference value, i.e., the population response when presenting a stimulus. It shows that the population distribution systematically encodes the value of the stimulus, but that the distribution is skewed for extreme stimulus values, which complicates the decoding process.

(Aldrich, 1997). After the model was trained, the histogram of the FSA response given to each stimulus was recorded. Then in the test session, a stimulus was presented, the FSA responses were read out, and the most likely stimulus that would give rise to this pattern of responses in the recorded patterns was located. Because this model did not involve noise, this approach can always give perfect results using the response histogram comparison.

However, this approach requires full information of the responses to all stimuli, and it is an indirect way to map the response to a discrete category. For adaptation, it is needed to compute a continuous decoded value instead. So, when estimating an adapted response, the two most likely recorded responses (using minimum mean square error, MSE) were chosen and the perception ρ was represented as the linear distance between their corresponding stimuli, calculated as

$$\rho = s_1 + \frac{MSE_1}{MSE_1 + MSE_2} \times (s_2 - s_1), \tag{3.6}$$

where s_1 and s_2 stand for the most likely and second likely stimulus, and MSE_1 and

MSE_2 stand for the mean square error of the most likely and the second likely recorded responses. Though we consider this approach a reasonable mechanism for computing a continuous value, it is only a linear interpolation, not a probably optimal value.

Another approach is support vector regression (SVR; Cristianini and Shawe-Taylor, 2000). This method is trained on all FSA activity patterns and their associated test stimuli, and computes a continuous hyperplane that fits these patterns. Since the hyperplane is continuous, it is able to fit unseen data. In the case of a new pattern with a slight difference from the trained one (e.g., in the case of adaptation), this approach is reliable and straightforward. SVR also has clear weakness: it is a complex method to decode stimuli, and how a brain may actually implement it is hard to tell.

In any case, determining how the rest of the brain may interpret activity patterns in the ventral stream is an unsolved problem, and here a reliable method for understanding and quantifying the behaviour of the model network was mainly used. Figure 3.6 shows the estimates of each stimulus using the population vector (a), ML (b) and SVR (c) against the veridical estimate using the same scenario as above (train on 10 caricaturisations, and test on 37 different ones). It can be seen that the population vector cannot capture the correct value at extremes, while the interpolated ML and SVR largely follow the veridical value, with SVR deviating at the most extreme values. Thus ML and SVR are better candidates to be used in simulating aftereffects.

However, there are two serious limitations for the ML and SVR methods applied to this model, which were thus implemented only for the one-dimensional case. The main issue is practical: extending the implementation to represent the multi-dimensional face space would have taken significantly more development time, and moreover requires an impractical level of computational resources. Another issue is in terms of biological plausibility — e.g. the ML method used in this model assumes a normal distribution, and the SVR behaves as a classifier by searching for a hyperplane to approximate the response pattern's stimuli. These specific approaches could be hardwired, but they seem relatively complex in terms of neural implementation. One may

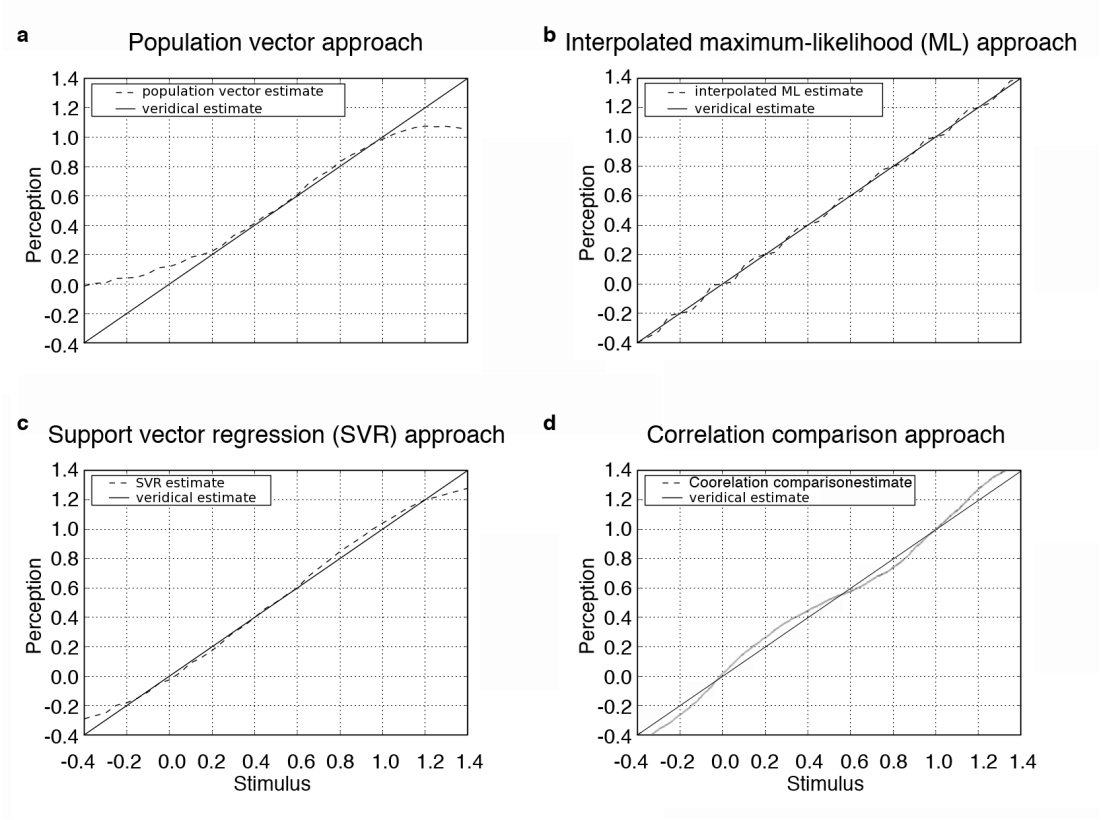


Figure 3.6: Estimated and veridical perception performed by four decoding methods before adaptation, based on the trained LISSOM model after 20,000 iterations. Perceptions are shown against one-dimensional face stimulus value.

even argue that the brain actually has no way of "decoding" — it is just matching responses with existing memory (e.g., Hawkins and Blakeslee, 2005 and George and Hawkins, 2005). Thus, we decided to develop a new decoding method to fit the two-dimensional face generator. Based on this idea, instead of training on the mapping between FSA activity pattern and face stimulus pattern, one can decode by comparing FSA responses directly. A correlation comparison decoding method was devised for this purpose. Consider $\mathbf{f_s}$ as the FSA activity pattern of stimulus s, and $\mathbf{f_0}$ as the FSA activity pattern of the pre-adaptation strength 1.0 (target) face. Pearson's correlation $\mathbf{c}_{\mathbf{f_s},\mathbf{f_0}}$ is defined as

$$\mathbf{c}_{\mathbf{f_s},\mathbf{f_0}} = \frac{\sum_{i=1}^{n} \left(\mathbf{f_s}_i - \bar{\mathbf{f_s}}\right)\left(\mathbf{f_0}_i - \bar{\mathbf{f_0}}\right)}{(n-1)\, s_{\mathbf{f_s}} s_{\mathbf{f_0}}}, \tag{3.7}$$

where $\mathbf{f_s}_i$ and $\mathbf{f_0}_i$ stand for the i^{th} element of vector $\mathbf{f_s}$ and $\mathbf{f_0}$, n for the total number of elements in these vectors, $\bar{\mathbf{f_s}}$ and $\bar{\mathbf{f_0}}$ for the mean of these vectors, and $s_{\mathbf{f_s}}$ and $s_{\mathbf{f_0}}$ for their standard deviation.

After adaptation, the FSA activity pattern for a stimulus has changed but the pattern of the pre-adaptation target face is intact, and this gives rise to the changed perception. This method is much more straightforward and intuitive than the interpolated ML and SVR methods. In addition, as it computes on the FSA activity pattern directly, it is independent of the stimulus space dimension. Decoding perception by this method for pre-adaptation stimulus yields results that were also very close to veridical perception (See Figure 3.6d).

3.2.4 Local decoding method for full face perception shift

The previous methods allow a LISSOM activity pattern to be decoded, but to truly understand how adaptation changes the representation of faces, it is necessary to determine the effect of adaptation at every point in face space. Although the face space is theoretically continuous, it may suffice to show discrete approximations of such changes. In the literature of face space theories, this pattern of changes is called "face

space shift". The correlation comparison method described in section 3.2.3 is based on the idea of "face memory", and should be more relevant to biological implementations than the much more complicated interpolated ML or SVR decoding methods. However, the constraint of correlation comparison is that it computes correlation only with the response of a particular stimulus, normally the target face. More intuitively, when the brain decodes faces, it should not be necessary to directly decode its coordinates in the face space (as did in ML and SVR trials), nor is it necessary to show the correlations with a particular stimulus as what was done when duplicating the Leopold et al. (2001) experiment.

A new and more general decoding method can be devised to show the change in the perception direction of a face after adaptation. Intuitively, when presenting a particular face, the human brain may first compare it with a similar face in memory in order to see if it can be recognised. This step can be very efficient and cost little energy. The paradigm of face space fits this approach — faces that are very similar are closer, and those that are not very similar are more distant in face space. Therefore, when presenting faces after adaptation, it is reasonable to suggest that the brain just searches in the neighbourhood of the presented face and performs recognition, because the neural activities elicited before and after adaptation by the same face are very similar to each other. In a typical adaptation task where faces have been in short-term memory, the brain does not need to perform a global search across the entire face space for recognition.

In the current model, this local comparison method was implemented using a discrete sum-of-vectors at each point of face space (see Figure 3.7 for an example). In the current two-dimensional face space, suppose the perception shift of face f_{mn} is being measured, where (m,n) denotes its location, i.e., height and size value in the face space. Define $d_{mn,ij}$ as the distance of the FSA activity between face f_{mn} and its neighbour f_{ij}, and H as the set of all the faces in the neighbourhood where f_{ij} resides. Then $d_{mn,ij}$ can be computed as

$$d_{mn,ij} = \sum_{f_{ij} \in H} \left(\mathbf{C}_{mn} - \mathbf{C}_{ij} \right)^2, \tag{3.8}$$

where $\mathbf{C}_{mn}$ and $\mathbf{C}_{ij}$ denote the FSA activity vector when presenting face f_{mn} and f_{ij}. Then the distance of face f_{mn} from all its neighbours can be obtained. If the neighbourhood radius to consider is set to 1, there are 8 $d_{mn,ij}$. The pre-adaptation distances $d'_{mn,ij}$ and post-adaptation distances $\hat{d}_{mn,ij}$ can be computed, and the shifted distance of face f_{mn} with its neighbour f_{ij} can be evaluated as

$$m_{mn,ij} = \hat{d}_{mn,ij} - d'_{mn,ij}. \tag{3.9}$$

Now the shifted perception of face f_{mn} after adaptation can be defined as the sum of the vector of its neighbouring faces' spatial difference in the face space weighted by their shifted distance $m_{mn,ij}$, as

$$\mathbf{S}_{mn} = -m_{mn,ij} \begin{bmatrix} i-m \\ j-n \end{bmatrix}. \tag{3.10}$$

Because positive $m_{mn,ij}$ means f_{mn} is more unlike f_{ij} (elongated distance) after adaptation, i.e., in the opposite direction to f_{ij}, and negative $m_{mn,ij}$ means f_{mn} is more like f_{ij} (shortened distance) after adaptation, i.e., in the same direction as f_{ij}, $m_{mn,ij}$ is made negative in the above equation to represent the flipped direction.

These definitions are illustrated in Figure 3.7, showing how the perception of a face can be computed as the sum of its neighbouring perceptions. In this chapter only the results for a two-dimensional face space were obtained, but this scheme can be naturally extended to higher dimensions using the same vector-sum-based scheme. However, the effect of face space shift may be hard to visualise beyond three dimensions.

3.2.5 Aftereffects simulation

The methods for building the model, generating face stimuli, and decoding model output have been discussed in previous sections. The face aftereffect simulation can

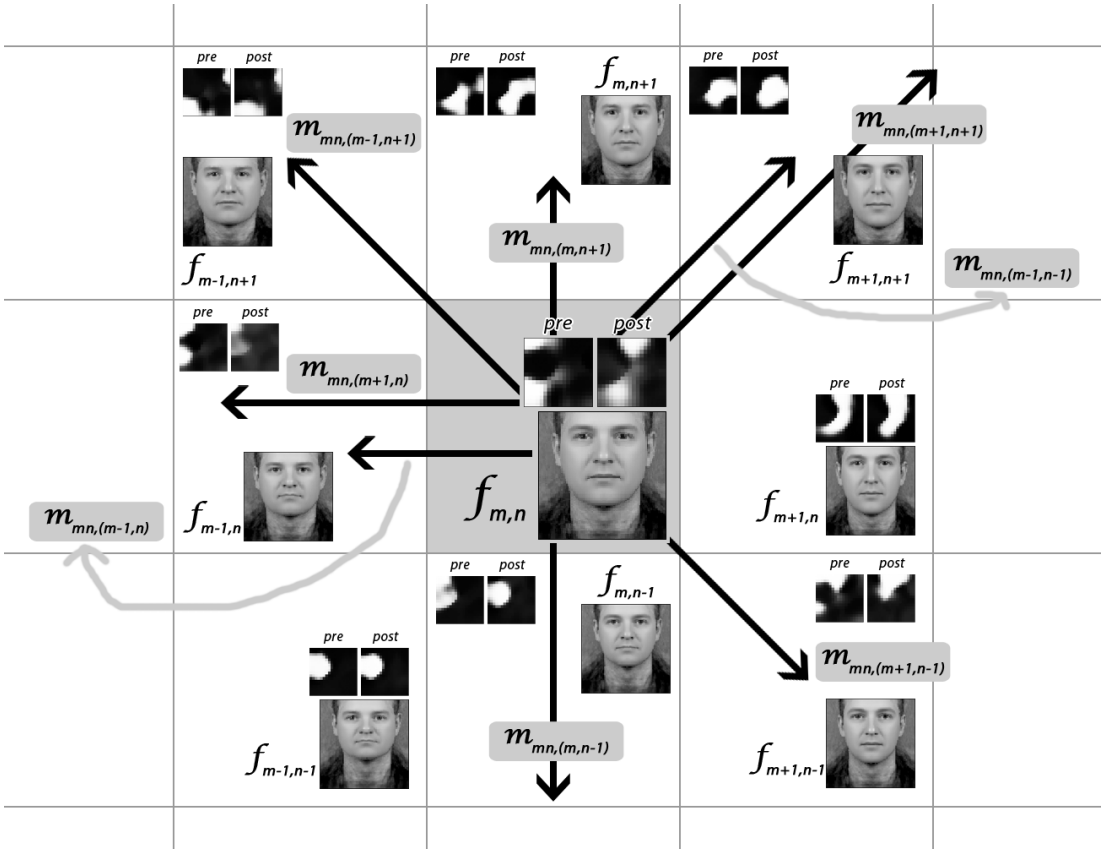


Figure 3.7: Illustration of indirect decoding of face space shift based on local comparison. The shifted perception of face f_{mn} is to be measured after adaptation. $m_{mn,(m-1,n-1)}$ etc. denotes the shifted distance of face f_{mn} with its neighbours. The arrows represent vectors with magnitudes $m_{mn,(m-1,n-1)}$, and their sum is $\mathbf{S}_{mn}$.

now be performed with these in hand.

An aftereffects simulation was first performed to duplicate the experiments in the work of Leopold et al. (2001), using the correlation comparison method described in section 3.2.3.

In pilot studies, results were comparable when either all projections (afferent, lateral excitatory, lateral inhibitory) had the same learning rate, or whether only inhibitory learning was allowed. This result is compatible with previous modelling results (Bednar and Miikkulainen, 2000), and with physiological data (Vidyasagar, 1990). Accordingly, to make the mechanism of aftereffects clear, only lateral inhibitory learning was used for these studies.

A complete process of the simmulated adaptation was performed in this way, using 20 iterations of adaptation. First, 20,000 iterations were performed to train the model with input generated by the two-dimensional (face height and size) face generator. The two face dimensions were drawn from a uniform random distribution from -6.0 to 6.0. In reality, human faces are more likely to follow a Gaussian distribution for most feature dimensions, but in this small two-dimensional face space, a uniform random distribution suffices to maximise dimensional variances and make the model more capable to distinguish different faces.

Then, a similar testing paradigm to the one used in the experiment conducted by Leopold et al. (2001) was followed. A particular face was chosen as a target face and its trajectory as one of the four faces to discriminate among. In the baseline session, the model was presented with strength from -0.2 to 1.0 target faces and their associative FSA activity patterns were recorded. In the matching adaptation session, the trained model was exposed to an anti-target-face (i.e., strength -1.0) for 20 iterations, and then the stimuli were presented and FSA patterns were recorded the same way as in the baseline session. In the non-matching adaptation session, the trained model was exposed to the anti-face of another visually distinctive trajectory for 20 iterations, and then the stimuli were presented and FSA patterns were recorded the same way as in

the baseline session. As mentioned above, in both adaptation sessions, only lateral inhibitory learning was enabled. The test faces, matching adaptation anti-face and non-matching anti-face, together with their associative FSA activity patterns are shown in Figure 3.8. The quantitative results and comparison with Leopold et al. (2001) is presented in Figure 3.9.

In order to understand the model mechanisms more fully, the indirect local comparison decoding method described in Section 3.2.4 will also be used to represent the perception shift. Unlike in the human experiment where only two conditions were performed (i.e., matching and non-matching anti-face, as shown above) during simulation, all possible adaptation conditions for all stimuli of a stimulus space will be simulated.

To clarify how the method works in a simple case, results will first be presented for decoding a two-dimensional position (x,y) aftereffect (previously described in Section 2.3). The stimuli in this case were vertical ellipsoidal Gaussian bars at different positions. The test position ranged from -0.75 to 0.75 at the step of 0.15 in both the x and y-axis, and the adaptation positions were also from -0.75 to 0.75 at the step of 0.15 in both the x and y-axis. Thus, in total there were 121 adaptation conditions and in each condition the perception changes of 121 positions were measured (presented in Section 3.3.2).

Second, results for adaptation in the two-dimensional face space will be presented (Section 3.3.3). The simulation process was the same as the position aftereffect, with adaptation and test face strengths ranging from -1.0 to 1.0 at the step of 0.2, leading to 11 values for each dimension and 121 different faces in total for adaptation.

3.3 Results and discussion

In this section, the modelling result for duplicating the experiment conducted by Leopold et al. (2001) using the correlation comparison decoding method, as well as the results for simulating the full perception shifts of the two-dimensional position aftereffect and

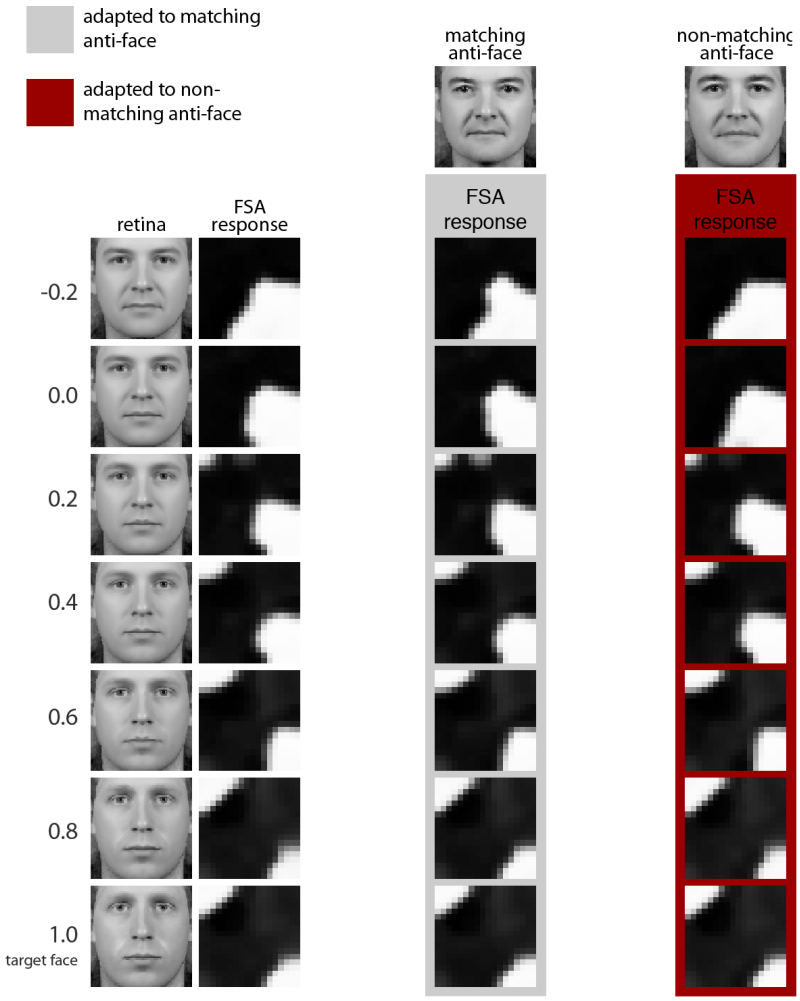


Figure 3.8: Paradigm of adaptation simulation. During adaptation, the lateral inhibitory learning rate was set to 0.095 (obtained from empirical pilot experiments) and all of the other learning rates were set to 0.0. The adaptation period was 20 iterations. Adaptation for all cases were performed on the LISSOM model trained for 20,000 iterations for each condition baseline before training (left column), matching anti-face (middle column) and non-matching anti-face (right column). The FSA response to each test pattern is shown with the response to the adaptation pattern shown at the top. Before adaptation, responses vary systematically for each face in the dimension (shown in the left column). After adaptation to face with strength -1.0 at the same trajectory as the target 1.0 face in the left column (i.e., the top face in the middle column), the response to each face has changed due to lateral inhibitory only learning during adaptation. After adaptation to a face with strength -1.0 at a different trajectory with the target 1.0 face (i.e., the top face in the right column), the response to each face has also changed but in a different way. This is caused by different adapting faces. By comparing the FSA responses visually to the baseline values, on can get an intuitive feel for the shifts in perception, which are quantified in Figure 3.9.

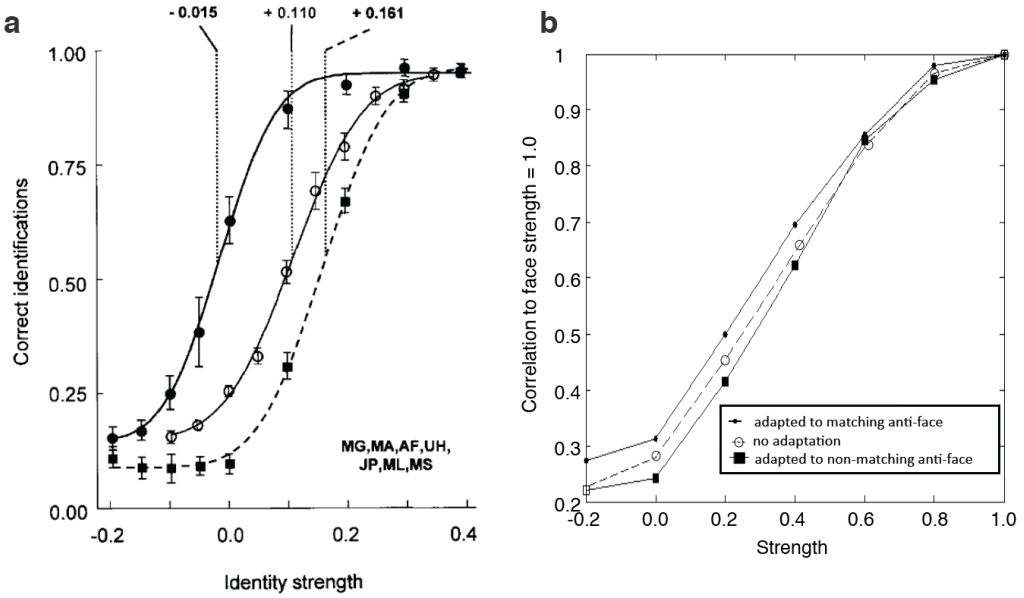


Figure 3.9: Psychophysical and modelling results for face perception shift. **(a)** Average human performance under adaptation conditions: no adaptation (open circle), adapting to matching anti-face (closed circle) and adapting to non-matching anti-face (closed square). **(b)** Decoded FSA activity pattern using correlation comparison under three adaptation conditions: no adaptation (open circle dashed line), adapting to matching anti-face (closed circle) and adapting to non-matching anti-face (closed square).

face identity aftereffects, will be shown and discussed.

3.3.1 Duplicating Leopold et al. (2001)

Using the decoding method described in section 3.2.3 and the simulation process from section 3.2.5, the main result of the experiment conducted by Leopold et al. (2001), i.e., the perception shift under opposite adaptation conditions, was duplicated. This result is shown in Figure 3.9b and aligns with the experimental psychometric curve in Figure 3.9a. It can be seen from the figures that the modelled result can be qualitatively compared to the psychophysical result as they show a similar trend of perception curve shifts under two opposite adaptation conditions.

Although the above result shows duplication of the experimental result, the methods (LISSOM model and correlation comparison decoding method) provided only one

possible theory to account for the underlying neural mechanism. Importantly, the choice of non-matching anti-face in the model is arbitrary, and thus the underlying criterion for choosing a "suitable" non-matching anti-face is poorly defined. Understanding what kind of stimulus can elicit a leftwards or rightwards perception shift is essential to help understand the "big picture" of the neural mechanism of face adaptation. Thus it is useful to learn about the systematic mechanism underlying the shift of the entire face space under all adaptation conditions.

3.3.2 Perception shifts of two-dimensional position aftereffect

As mentioned previously, to see if the indirect local comparison decoding method works properly, it is helpful to test it on a known low-level aftereffect. In this case, a suitable choice is a two-dimensional position aftereffect. Following the simulation process mentioned in section 3.2.5, the result is shown in Figure 3.10.

Each subplot in Figure 3.10 stands for an adaptation condition. I.e., after being adapted to a position denoted as the location of this subplot in the entire figure, how does the perception of all the faces change? The direction of arrows in a subplot stands for the direction of the perception shift of the Gaussian bar at that position under this adaptation condition. It can be seen that apart from the subplots at the outermost locations (i.e., subplots with x or y values of ± 0.75), the perception changes around adaptation points are repulsive after adaptation, and nearby points are perceived as being further away than they really are. This result matches previous findings for the TAE, as discussed above. Responses for stimuli at the edges are difficult to interpret, as they represent neurons at the edges where the network does not organise well.

Thus the indirect decoding scheme shows a repulsive shift in perception around the adaptation point for the position aftereffect with the magnitude of perception changes peak in around the adaptation point. This result appears consistent with the shifts in perceived position following adaptation to visual motion (Snowden, 1998), as well as those in one-dimensional TAE. Thus, this method appears to offer a good way to

−0.75
−0.6
−0.45
−0.3
−0.15
0
0.15
0.3
0.45
0.6
0.75
0.75
0.6
0.45
0.3
0.15
0
−0.15
−0.3
−0.45
−0.6
−0.75

Figure 3.10: (previous page) The LISSOM network adapted to a vertical line (0 degree) at 121 locations of the retina. Values in the x- and y-axis (range: -0.75 to +0.75) represent locations in the retina. Each subplot represents the perception changes of test stimuli after adaptation to the stimuli at corresponding retina location. In each adaptation condition (subplot), perception changes were measured by the local-comparison based indirect decoding scheme for the test stimuli at the same 121 locations. The results are shown as 121 subplots, where each subplot shows that adaptation affects stimuli in a small neighbourhood around the coordinates of the adaptation stimulus, with the perceived value typically shifting away from that stimulus.

understand and visualise how the perceptual space varies with adaptation.

3.3.3 Perception shifts of two-dimensional face aftereffect

The results for the two-dimensional face aftereffect are shown in Figures 3.11 and 3.12. The meaning for these results is exactly the same as that described in the previous section and in Figure 3.10, except that each subplot and each arrow in a subplot represents a face rather than a position.

In the pilot study, it was observed that for TAE and position aftereffects, the size of a neighbourhood can affect the organisation of perception changes. Two neighbourhood sizes, 2 and 6, are plotted as examples here. These are shown respectively in Figure 3.11 and 3.12. It can be seen that the bigger the neighbourhood, the more organised the perception change. In the plots for neighbourhood size 6, arrows exhibit a regular radial manner starting from each of their centres of gravity; while in the plots for neighbourhood sizes 2, this is less regular. Compared with Figure 3.11, Figure 3.12 shows a more consistent result with TAE and the position results, effectively averaging over more values to make the overall pattern of adaptation clearer.

For both neighbourhood sizes, it can be seen that the perception changes are highly correlated with the adaptation points, and the changes around the adaptation points are repulsive overall. As in the position aftereffects, the magnitude of perception changes

−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

Figure 3.11: (previous page) The LISSOM network adapted to face stimuli whose height and size value are determined by the x- and y-axis values (range: -1.0 to +1.0). Each subplot represents perception changes of test faces after adaptation to the face of the corresponding x (height) and y (size) value. There are in total $11 \times 11 = 121$ adaptation conditions. In each adaptation condition (subplot), perception changes were measured by the local-comparison based indirect decoding scheme for the test faces of the same 121 values. The results are shown as 121 subplots. Neighbourhood size: 2. Though the pattern is less regular than in Figure 3.10, changes are typically repulsive and are centred around the adapted location in face space.

peaks around the adaptation point. It can be seen that in the case of face aftereffects, such a correlation and repelling pattern is not as clearly organised as in the position aftereffect case. This result is likely to be due to the non-independency of dimensions in the face space, as discussed previously, which is realistic for this more complicated perceptual space. Even so, these results mean that, measured by the indirect local comparison decoding method, the face aftereffects elicit a perception change qualitatively similar to low-level effects like position aftereffects and the TAE. This strongly suggests that the theory for low-level effects can be used consistently to explain face aftereffects, at least in the current two-dimensional face space. The modelling work presented in this chapter could thus provide a theoretical link between neural mechanisms for low-level and high-level visual aftereffects.

To establish a clear link to neural mechanisms, future experimental work will be needed. The above results are currently limited to illustrating the perception changes around the adaptation point. Yet the results from the experiment conducted by Leopold et al. (2001) focused on the perceptions at the opposite side of the adaptation face, which, in the above modelling results, are changed relatively little compared to the changes around the adaptation point. The modelling using mechanisms from low-level perception thus predicts that (a) changes are repulsive around the adaptation stimulus,

−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

Figure 3.12: (previous page) Same network and adaptation condition as Figure 3.11, but with decoding neighbourhood size 6. Due to the greater averaging, results are clearer and more regular with this large neighbourhood size.

and (b) changes reach greatest values at some point near the adaptation point in perceptual space. The results suggest testing these predictions for the perception shifts around adaptation point psychophysically, to revoke or validate the proposed model. However, it is difficult. For practical reasons, evaluating a two- or multi-dimensional face space in a human experiment is not practical as they typically require a prohibitively long experiment due to the large number of test faces.

Instead, one could investigate a single dimensional effect, systematically investigating values near and distant from the adaptation point. In the next chapter, modelling work based on these ideas will first be performed, and in Chapter 5, human experiments based on a similar idea will be conducted.

3.4 Conclusion

This chapter described in detail how the face identity aftereffect was modelled. The LISSOM architecture was first introduced. Then one-, two- and three-dimensional face generators were explained in detail. To make the modelling work feasible and plausible at the same time, the two-dimensional shape-only face generator was finally chosen in the aftereffect simulation. Four types of decoding method — population vector, interpolated maximum likelihood, support vector regression and correlation comparison, were discussed, and the correlation comparison method was finally chosen for the aftereffect simulation. With the above-mentioned approaches, the main results for face identity aftereffect described in the experiment conducted by Leopold et al. (2001) were duplicated.

While Leopold et al. (2001) and the later experimental literature are not able to

show the big picture of how every face in a face space could shift under all possible adaptation conditions, computational models can help to achieve this. Understanding such a big picture is very helpful to reveal the underlying neural mechanism of adaptation. For this purpose, an indirect local comparison method was devised and the full perception shifts for both low-level position aftereffect and high-level face identity aftereffect were computed. These results showed similar perception shift trends and strongly suggested that low-level and high-level adaptation may be explained in a unified underlying neural theory. Interestingly, the analysis of these models predicts a clear pattern of repulsive effects around the adaptaion stimulus, and moreover that the effects should be largest near the adaptation stimulus. Existing experiments have not tested these predictions.

To investigate this theory, a revised adaptation paradigm and simplified face space (one-dimensional) will be used for modelling in the next chapter, and then tested using human experiments in the following chapter, thus completing a loop from the initial experimental results of face aftereffects, to computational modelling to understand possible mechanisms, and then experimental tests of predictions from these models.

Chapter 4

Modelling one-dimensional low- and high-level aftereffects

This chapter first briefly reviews previous experimental work on one-dimensional low-level and high-level aftereffects. Then, modelling of one-dimensional tilt aftereffects (TAEs) and face gender aftereffects (FAEs) is done using two complementary approaches — a LISSOM-based network and a simplified exemplar-based multichannel model. Last, predictions obtained from this modelling work are discussed, which lead to the experimental work described in Chapter 5.

4.1 Introduction

The modelling results from Chapter 3 suggested that neural adaptation may be similar between high-level (FAE) and low-level (TAE) effects. It is therefore essential to test this theory experimentally. However, it is hard to apply the same paradigm used in the modelling in Chapter 3 for human psychophysical experiments, because:

1. Face space with two or more dimensions is impractical to test in human experiments. For example, the stimulus space shown in Figure 3.11 and Figure 3.12 involves at least 121 face stimuli. In a typical psychophysical experiment, each

stimulus is tested in repeated trials in order to verify that the results are reliable. Moreover, it is important to test a large number of faces in the stimulus space in order to obtain accurate measurements. For a psychophysical experiment involving adaptation in each trial, such a paradigm would require several hours of continuous experiments. Such long experiments would raise ethical and health issues for human participants, and moreover the results will be vastly unreliable due to fatigue and loss of concentration.

2. Previous experimental paradigms are not suitable for conducting comparable experiments. For example, when adapting to a face far from the average face, e.g., one condition in the upper-left corner in Figure 3.11, the experimental participant will be tested for a perception shift around that adaptation face. However, the faces around that adaptation face are all far away from the average, and therefore there is no natural boundary among them like face gender. In this case, one can only use subjective rating such as "masculinity" to measure the perception shifts around the adaptation face. This psychophysical approach is not always reliable and may lead to unknown results.

In order to test the theory proposed in the previous chapter, this chapter focuses on a one-dimensional stimulus space, which makes both experimental and modelling work much more feasible. As reviewed in Section 2.3, many low-level one-dimensional aftereffects exhibit an S-shaped aftereffect curve that is qualitatively similar to the TAE, such as position (following adaptation to visual motion) and size aftereffects. However, as reviewed in Section 2.2, no studies have reported this property in clearly high-level one-dimensional aftereffects such as face gender, face race or face identity aftereffects.

In this chapter, both one-dimensional TAE and FAE models will be constructed, distilling the predictions of the model down so that they are easily testable. First, the mechanistic LISSOM-based network will be adapted for a one-dimensional TAE and

FAE. Then, a more abstract exemplar-based multichannel model will show how the basic principles can be exhibited with only a few general assumptions. In each case, the models will predict that aftereffect strength will have an S-shaped curve, a key prediction that will be tested experimentally in Chapter 5.

4.2 Modelling one-dimensional TAE and FAE: LISSOM-based network

In this section, the details of modelling the one-dimensional TAE and gender FAE based on the LISSOM network are described. The model architecture, adaptation simulation process and decoding method are explained only briefly, as they are the same as those described in Section 3.2. This section focuses on the different one-dimensional stimulus space and results. As discussed above, the purpose of the work presented in this section is to show that the same model that gave rise to the turning point of the aftereffect values for a two-dimensional stimulus space will also produce a similar curve in one-dimensional stimulus space.

4.2.1 Model architectures and stimuli

The model architecture used to simulate the gender FAE was exactly the same as was described in Section 3.2.1, including the position of the sheets and the size of afferent, lateral inhibitory and lateral excitatory connections. The architecture for the TAE was similar, but different in terms of the size of the V1 afferent and lateral connections. As shown in Figure 4.1, the size of the receptive field from V1 to LGNs and the size of the lateral connections for the TAE model are both much smaller than that in Figure 3.1. The radius of the V1 receptive field in this model equals 0.4 times that of the FAE, and the radius of the V1 lateral inhibitory and excitatory connections are 0.2 times that of the FAE respectively. These parameter values are similar to those of the previous

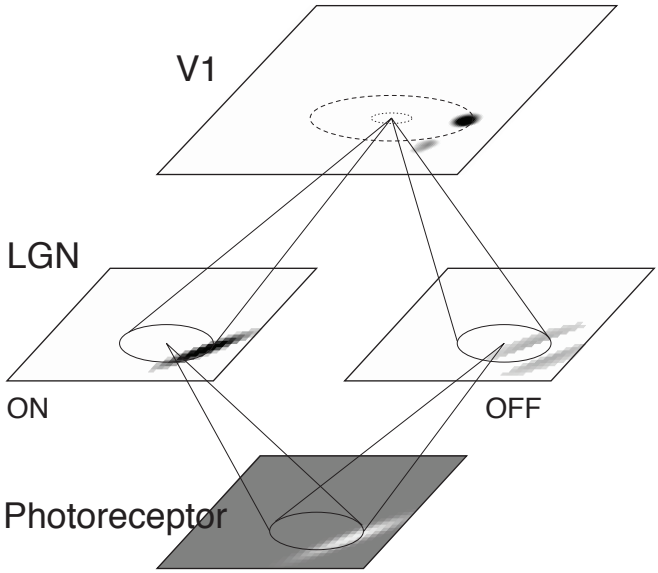


Figure 4.1: Diagram of a basic LISSOM network. LISSOM consists of a hierarchy of two-dimensional sheets of neural units, including an array of photoreceptors (**bottom**), ON and OFF channels in the LGN (**middle**), and a cortical network representing V1 (**top**). The photoreceptors can contain either oriented line segments as shown for the TAE, or faces for the FAE. This figure is from Miikkulainen et al. (2005)

LISSOM model on TAE (Bednar and Miikkulainen, 2000).

The stimuli for the TAE model are oriented Gaussian bars, because line orientation is π-periodic ranging from -90° to 90° (i.e., -90° and 90° stimuli are visually identical). As illustrated in Figure 4.1, during training they were presented on the photoreceptor sheet in a random location, in order to produce a topographic map. Samples of these stimuli are shown in Figure 4.2. In the test stage, stimuli were presented at the center of the retina.

The stimuli for the FAE model are a set of gendered face continuum examples, from extremely masculine to extremely feminine faces. There were 501 composite faces generated using the methods described by Tiddeman et al. (2001) from the 150 young adult male and 150 young adult female Caucasian photographic faces collected by Ian Penton-Voak at Stirling University (Penton-Voak et al., 2006). For each photograph, Penton-Voak et al. manually marked key locations (173 points) around the main

a -60˚ and 10˚ **b** -50˚ and -70˚ **c** 30˚ and 80˚

Figure 4.2: Sample Gaussian bar stimuli plotted on the retina, used in the TAE model. Two stimuli at different orientations were presented simultaneously in each iteration. Angles are relative to the vertical (0°), and they increase counterclockwise.

features and the outline of the face. The photographic quality results of an average male, average female and an average of the whole population (denoted as -0.5, +0.5 and 0 morphing strength) were created with the average value of the features across all males, females and the whole population, respectively. These average faces were generated by triangulating all of the key locations, computing average shape vectors from these locations, warping all of the faces into that shape, and computing average facial images. Principal component analysis (PCA) was performed on both the shape-free face images and the image-free face shape vectors to produce eigenfaces and eigenshapes. Taking these average faces as a basis, a face can be reconstructed by adding eigenfaces to the average shape-free face and then distorting it by applying eigenshapes. Thus, a face morph between two average faces can be generated. The 501 discrete faces used in this study were generated by linear combinations of these eigenfaces and eigenshapes. The three average faces were key frames in a face continuum covering a morphing strength from -2 to +2, generating a wider variety of face shapes than the -1 to +1 range used previously. A sample of the generated synthesised face continuum is shown in **Figure 4.3**. The other technical details for how to generate this set of faces are exactly the same as those described in Section 3.2.2.

Figure 4.3: Example face stimuli used in the FAE model, covering a wide range of male and female faces. 13 examples out of the 501 faces making up the continuum are shown, from the most masculine (-2.0) to the most feminine (2.0) face.

4.2.2 Decoding method and aftereffects simulation

The indirect local comparison decoding method was used to decode both the TAE and FAE in this chapter, in order to compare them with the results of the two-dimensional aftereffects decoded by the same method. The details of this method can be seen in Section 3.2.4. Note that the major difference between the TAE model shown in this chapter and the one described in Bednar and Miikkulainen (2000) is the decoding approach. Likewise, the process of aftereffects simulation is also identical to that explained in Section 3.2.5.

Note that the TAE measurement described in this chapter was based on one TAE simulation with a fixed position of adapting stimulus, instead of averaging across many simulations with random position of adapting stimulus as in experimental studies and previous modelling studies. As can be seen in Figure 4.4b, the resulting curve is less smooth and less symmetrical than in the previous model (Figure 4.4c), but it still maintains the overall shape. Future work can generate smoother and more symmetric curves by running these additional randomisations.

4.2.3 Results

Following the methods described above, the results for modelling the one-dimensional TAE are shown in Figure 4.4a,b. The LISSOM network was first trained for 20,000 iterations by random Gaussian bars at random positions in the photoreceptor sheet. Then the LISSOM network adapted to a Gaussian bar at a particular orientation (i.e., one adaptation condition) at the centre of the photoreceptor sheet for 90 iterations. The

TAE was measured for test lines at each orientation by subtracting the pre-adaptation perceived orientation from the post-adaptation perceived orientation. Therefore, a positive TAE value means an increased orientation difference after adaptation (perceived orientation has moved length away from the adaptation orientation), and negative TAE means a decreased orientation difference after adaptation.

Figure 4.4a shows the result for seven adaptation conditions of TAE based on a local comparison decoding method, with each row representing an adaptation condition. The middle row corresponds to the vertical line (0°) adaptation condition. A left arrow denotes a decreased perceived angle value and the right arrow denotes an increased angle value. The same data can be used to construct the continuous aftereffect curve in Figure 4.4b.

Figure 4.4c shows how this result can be compared with previous LISSOM network results decoded by the population vector sum (Bednar and Miikkulainen, 2000), and Figure 4.4d shows how it can be compared to the experimental results of Mitchell and Muir (1976). It can be seen that the central portion of Figure 4.4b is slightly anti-symmetric around 0°, compared to Figure 4.4c,d, and at both ends of the curve, i.e., towards $\pm 90^\circ$, the magnitude of the aftereffect is larger than in the population vector sum model and the human experiment. These differences are within expected values, as the curve shown is for a single adaptation rather than the averages used for the previous modelling and experimental work. The significant characteristics of the TAE curve are consistent for all three cases, i.e., the amount of aftereffect reaches a peak at around $\pm 10^\circ$ orientation difference, and soon decreases to zero or even changes sign thereafter.

In addition, it can be seen from other rows in Figure 4.4a that the aftereffect curves for those conditions are just like shifting the curve in Figure 4.4b leftwards or rightwards. This property is also consistent with existing psychophysical results (Mitchell and Muir, 1976).

It should be noted that although the shape of the curves in Figure 4.4b and 4.4c are

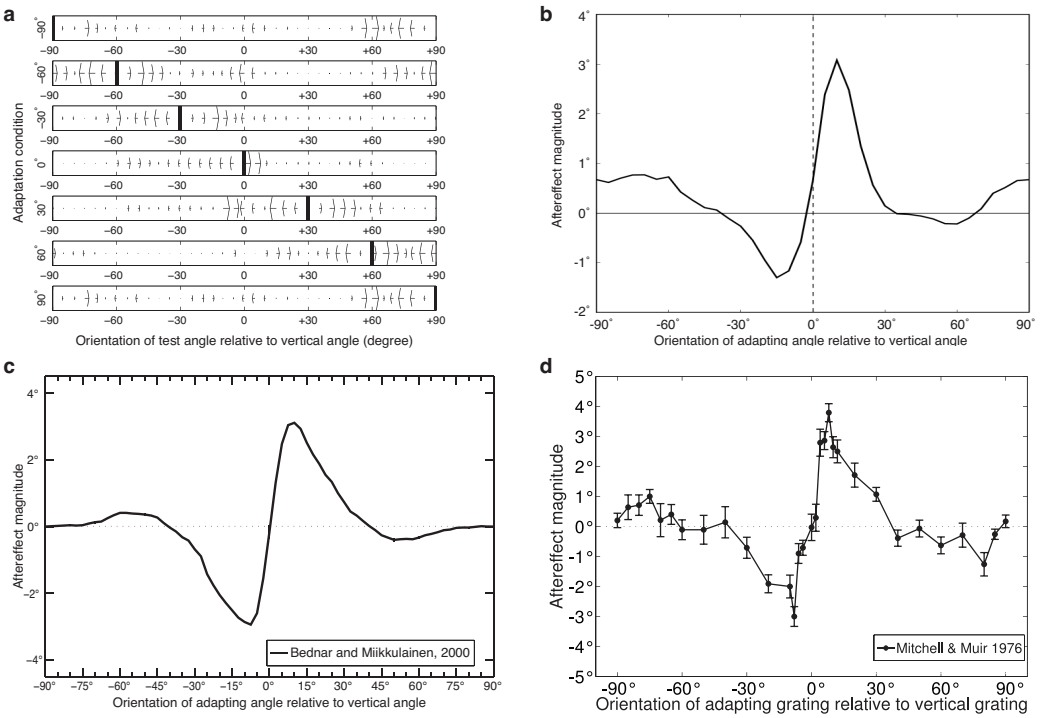


Figure 4.4: TAE perception shift and curves for LISSOM model and human data. **a**: Perception shift for the TAE in LISSOM under seven adaptation conditions, using the local comparison decoding method. Each arrow in the plot stands for the amount and direction of perception shift after adaptation to the corresponding adaptation Gaussian bar and the vertical black bar marks the adaptation line. The arrow size was computed by subtracting the pre-adaptation perceived orientation from the post-adaptation perceived orientation. **b**: TAE curve by the local comparison decoding method for the LISSOM-based network. This curve corresponds to the adaptation to a vertical line (0°) condition in the middle row of (a), plotted using the same data as for the arrows and connecting adjacent values. **c**: TAE curve by the population vector sum decoding method for the LISSOM-based network averaging over then trials. The current results are consistent with the previous decoding method. This figure is from Bednar and Miikkulainen (2000). **d**: TAE curve measured in the human psychophysical experiment by Mitchell and Muir (1976), which is consistent with both model decoding methods.

similar, the former is less smooth and less symmetric. As mentioned in Section 4.2.2, these issues can be rectified by averaging multiple simulations. Even so, the results should still hold, as Figure 4.4b has the same overall shape as in 4.4c.

Overall, the local comparison decoding scheme can well explain the TAE, and how this method can account for the more complex face gender aftereffects will be seen next.

Following the methods described above, the results for modelling a one-dimensional FAE for face gender are shown in Figure 4.5a,b. The LISSOM network was first trained for 20,000 iterations with uniform random faces drawn from the face continuum described in Section 4.2.1. Unlike in the TAE models, the face positions were fixed and aligned at the centre of their eyes during training and adaptation, to allow a small network to be used as representing only face shape and not location (ref. work by Bednar (2002) as an example of doing this). Then the LISSOM network adapted to one of various adaptation conditions (including the average (androgynous) face) for 250 iterations. Then FAE was measured for the test faces at all morphing strengths by subtracting the pre-adaptation perceived strength from the post-adaptation perceived strength. Therefore, a positive FAE value means the perception shifted towards more feminine faces, and a negative FAE means the perception shifted towards more masculine faces.

Figure 4.5a shows the result for five adaptation conditions of FAE based on the local comparison decoding method, with each row representing an adaptation condition. The middle row corresponds to the average (androgynous) face adaptation condition. A left arrow denotes a masculine shift and a right arrow denotes a feminine shift. Like in TAE, the discrete arrow can be computed continuously to form the aftereffect curve in Figure 4.5b.

A common feature of the previous TAE results is that for all adaptation conditions, the perception of a stimulus around the adaptation points is repelled away from the adaptor. Similar results hold in the case of face gender, as shown in Figure 4.5a,b.

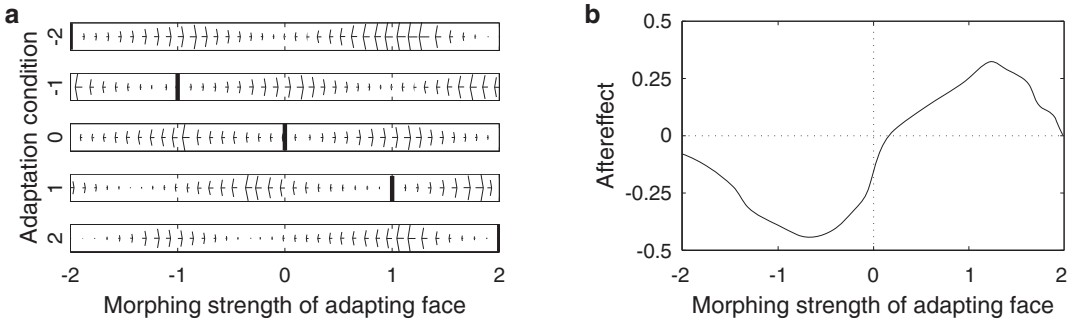


Figure 4.5: FAE perception shift and curves represented by the local comparison decoding method (LISSOM-based network). **a**: Perception shift of FAE under five adaptation conditions by the local comparison decoding method. Results were produced by the local comparison decoding scheme on the trained network. Each arrow in the plot stands for the amount and direction of the perception shift after adaptation to the corresponding adaptation morphing strength. They were computed by subtracting the pre-adaptation perceived morphing strength from the post-adaptation perceived morphing strength. **b**: FAE curve by the local comparison decoding method for the LISSOM-based network. This curve corresponds to adaptation to an androgynous face (morphing strength 0) condition in the middle row of (a), but similar results hold regardless of adaptation location.

Figure 4.4 and Figure 4.5 show that the modelled TAE and FAE effects are strikingly similar. The LISSOM-based network, together with the local comparison decoding method, thus suggests a similar trend of post-adaptation perception shift for both two-dimensional and one-dimensional tilt and face stimulus space, and thus suggests potentially similar underlying neural mechanisms.

4.3 Modelling one-dimensional TAE and FAE: simplified approach

Numerous computational and theoretical models (e.g., Bednar and Miikkulainen, 2000; Seriès et al., 2009, as well as the above-mentioned model) have been proposed to explain the TAE curve and similar low-level effects, but nearly all are based on three main principles:

1. visual cortex neurons only respond to a limited range of stimulus values (e.g., V1 neurons have a limited tuning bandwidth for orientation);

2. neurons activated by a stimulus adapt, reducing their responsiveness by some means (whether by increased inhibition as in LISSOM or by depletion of some resource; reviewed in Kohn, 2007), and

3. the perception at any instant is determined by the activity pattern across the population of neurons, such that the perceived value differs when some of the neurons are less responsive (see Seriès et al., 2009).

The modelling work for both TAE and FAE described in the previous section followed these three principles by showing how neural adaptation could arise from laterally inhibitory Hebbian learning, while showing how neurons could construct a representation for a multi-dimensional stimulus space allowing long-term development. However, the resulting model is complex and difficult to analyse, and it is possible to

formulate a simpler non-developmental mathematically tractable abstract model that can still account for TAE and give rise to the similar FAE and TAE curve. In this section, a much-simplified model is proposed to illustrate the theory of similar low-level and high-level adaptation mechanisms.

Figure 4.6a shows a concrete implementation of these principles using the motion-aftereffect model from Seriès et al. (2009), fit to the published TAE data. The model illustrates how neural adaptation leads to aftereffects with a realistic S-shaped curve (red curve in **Figure 4.6b**).

This simple TAE model was adapted from the model for the motion direction aftereffect (DAE) described by Seriès et al. (2009). This model consists of a population of neurons, each with a tuning curve most selective for a particular orientation. The population of N neurons with tuning curves

$$f(\theta) = \{f_1(\theta), f_2(\theta), \cdots, f_N(\theta)\} \tag{4.1}$$

describs the mean spike count of each neuron as a function of the stimulus direction θ. These N neurons were chosen to tile the space of all orientations uniformly and have unimodal tuning curves following the circular normal distribution (Seriès et al., 2009):

$$f_i(\theta) = G_i \exp\left(\sigma^{-1}(\cos(\theta - \theta_i) - 1)\right), \tag{4.2}$$

where the gain G_i controls the response amplitude of neuron i, θ_i its preferred orientation, and σ the width of each tuning curve. The encoding model was specified as the probability of observing a particular population response $r(\theta)$ for a given stimulus θ (see equation 5.2 in Seriès et al., 2009).

The biggest difference in this model from the LISSOM-based network described previously is the adaptation paradigm. When one orientation (the adaptor) is presented repeatedly, rather than adaptation emerging from changes in specific lateral inhibitory connections, in this model, neurons responding to that orientation reduce their responsiveness directly (see the three examples in **Figure 4.6a left column**). The gain is reduced more for the most active neurons and is based on a normal function of the

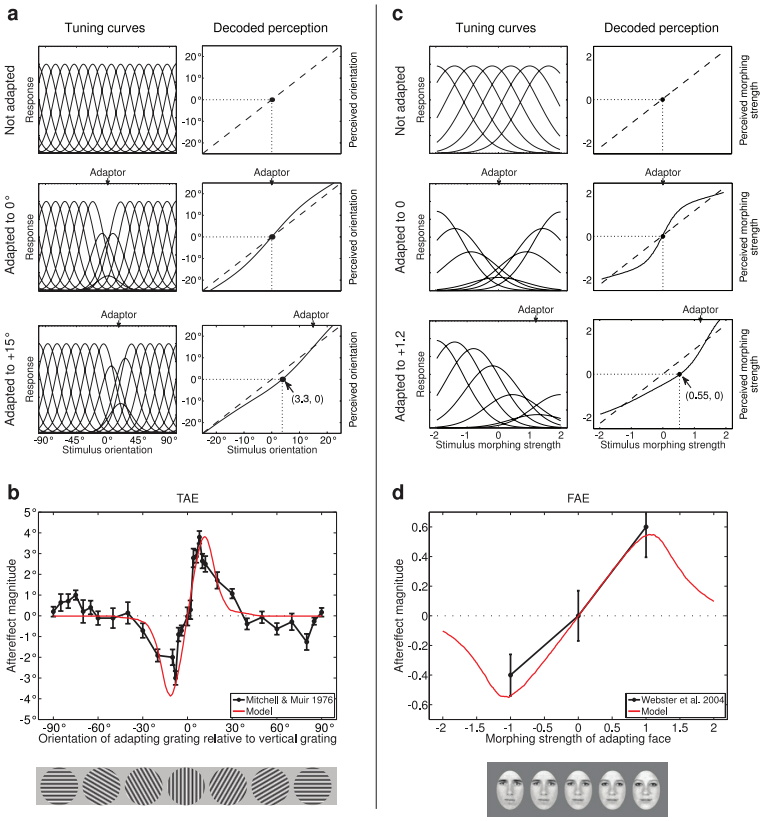


Figure 4.6: General model for aftereffects based on adaptation of neurons tuned to specific feature values. **(a, left)** Gaussian tuning curves for 15 neurons in a computational model of the TAE (adapted from ref. Seriès et al., 2009). Repeated presentation of one orientation (marked "Adaptor") reduces the responsiveness of neurons whose tuning curves overlap with the adaptor (i.e., neurons that responded to the adaptor). **(a, right)** Perceived orientation for each possible test pattern before (dashed line) and after (solid line) adaptation, calculated using a Maximum-Likelihood (ML) method. Filled circles show the new perceived "vertical" ($0°$). E.g., after adaptation to $+15°$, $+3.3°$ is now decoded as $0°$, yielding an aftereffect of $+3.3°$. **(b)** The red TAE curve summarises these shifts in preferred orientation for each adaptor, yielding a prediction of the tilt aftereffect strength comparable to the psychophysical results (e.g. from participant DEM of Mitchell and Muir, 1976 as shown here). **(c, left)** The same model as in **(a)** and **(b)** applied to a facial gender aftereffect (FAE). **(c, right)** For each adaptation condition, the population response that is ML-decoded as androgynous (0) is shown with a filled circle. **(d)** The model FAE curve matches the sparse existing data (Webster et al., 2004), but strongly predicts that FAE values will decrease for larger morphing strength magnitudes (more feminine or more masculine faces).

difference between the adaptor direction and the preferred direction of that neuron (see equation 5.3 in Seriès et al., 2009). This TAE model was identical to the DAE model, as published by Seriès et al. (2009), except for the specific parameter values chosen to fit the TAE data: tuning curve width $\sigma = 0.09\pi$ (was 0.18π in DAE), maximal suppression $\alpha_a = 100$ (was 50 in DAE) and the spatial extent of the response suppression $\sigma_a = 0.06\pi$ (was 0.125π in DAE). Details of the suppression model parameters are shown in equation 5.3 of Seriès et al. (2009).

To decode the responses, the unaware maximum likelihood (ML) decoder was used (described in Section 2.3 of Seriès et al., 2009). To match the experimental protocol to be discussed in the next chapter, how the perception of 0° changes with adaptation was measured, as illustrated in **Figure 4.6a**. Based on the decoder, the model predicts how the perception of subsequent test patterns will differ as a result of the changes in responsiveness (see the three examples in **Figure 4.6a right column**). Other TAE models differ in details, but generally follow the three principles above.

If high-level perception also follows the above-mentioned three principles, then one would expect face aftereffects to follow an S-shaped curve as well. To illustrate this idea, the Seriès et al. (2009) model was again modified, primarily to account for the qualitatively different facial gender input dimension in order to simulate face gender aftereffect (FAE). Instead of using the circular normal distribution specified in equation (4.2) above, which is suitable for cyclic quantities like orientation and direction, the FAE model used a non-cyclic normal distribution:

$$f_i(s) = \frac{G_i}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(s-s_i)^2}{2\sigma^2}\right), \tag{4.3}$$

where s denotes the stimulus morphing strength, and s_i the i^{th} neuron's preferred morphing strength; G_i and σ have the same meaning as in equation (4.2). To match the FAE data, the FAE model (**Figure 4.6c**) used broadly tuned neurons covering much of the input space, with $\sigma = 0.8$, $\alpha_a = 24$ and $\sigma_a = 0.64$ (where each is specified in units of morphing strength). Note that for the FAE, the stimulus value range ($[-2,2]$

morphing strength) was different from that of TAE ($[-0.5\pi, 0.5\pi]$), and therefore the corresponding parameter values have different scaling.

Apart from the non-cyclic input dimension, broadly tuned neurons, and adaptation patterns ranging from masculine (morphing strength < 0) to feminine (> 0) (**Figure 4.6c**), other parameters are otherwise identical to the TAE model. To compute the FAE, how the perception of a face with morphing strength 0 changes with adaptation was measured, as illustrated in **Figure 4.6c**.

As **Figure 4.6d** illustrates, the modified model does predict an S-shaped curve for the FAE, with aftereffect magnitude decreasing for sufficiently gendered faces. This result is consistent with the results from Section 4.2, suggesting that the three principles mentioned in this section are sufficient to give rise to similar TAE and FAE curves.

4.4 Discussion and conclusion

The local comparison decoding method used in this chapter can be applied in any stimulus space in any dimensions, as long as after adaptation the cortical activity pattern evoked by a stimulus has changed slightly. Because this local comparison scheme assumes that adapted perception can be related to responses to neighbouring stimuli, if there are dramatic changes in activity patterns, this assumption may not hold. This limitation should not be an issue for the one-dimensional TAE and FAE discussed in this chapter, as the results have shown.

It should be noted that it is only the perception shifts of average (androgynous) face that can be measured experimentally. In Figure 4.5a, for each row, although the perception shifts for all morphing strengths are computed, only the five points at 0 morphing strength has experimental correspondence (as also illustrated in Figure 4.5b). In TAE, when testing the orientation of a bar, the participant can easily report the degree by adjusting a paralleling bar, which is accurate and reliable. But in the case of judging face masculinity, there is no such referencing object, and experiments by

subjective rating are prone to be inaccurate. Therefore, those test cases apart from 0 morphing strengths in Figure 4.5a have been hard to verify experimentally so far, and only the perceptual shift of perceptual boundary (i.e., an androgynous face) can be tested in a human experiment.

Two kinds of models for both one-dimensional TAE and FAE were described in this chapter. The LISSOM-based network was proposed to show how concrete neural mechanisms also driving development, plus a decoding method, can lead to low- and high-level aftereffects. The exemplar-based multichannel model was proposed to illustrate the essential theory in a simpler and more general manner based only on assumptions about neural coding and changes in responsiveness. Their results suggested that if the same type of neuronal organisation and learning scheme were used, the modelled FAE could show similar results to that of TAE. This prediction is consistent with the prediction for two-dimensional low- and high-level stimulus space explained in Chapter 3. Together, they lead to a potential theory that FAE may share similar underlying neural mechanisms with TAE.

However, this theory may appear contradictory to existing psychophysical studies of human face gender, identity and distortion aftereffects (Webster et al., 2004; Little et al., 2005; Leopold et al., 2001; Jeffery et al., 2010) and physiological studies of the FAE in monkeys (Leopold et al., 2006) which have instead reported aftereffects that monotonically increase in magnitude as the adaptation conditions goes away from an average face (see Section 2.2 for a review). Therefore, in the next chapter, a psychophysical experiment on perceptual boundary will be performed to test this theory directly to see if Figure 4.5b will describe the human results when they are tested over a large enough gender continuum.

Chapter 5

Psychophysical experiments on face gender aftereffects and tilt aftereffects

This chapter describes in detail two psychophysical experiments. The experimental work on tilt and face gender aftereffects is introduced, followed by the subjects, apparatus, stimuli, procedure and results for each experiment. Both experiments are then discussed at the end, along with theories that could explain the results.

5.1 Tilt aftereffects experiment

The modelling results in the preceding chapters predicted that if face-selective neurons have a variety of band-limited tuning preferences for face gender, then the FAE and the TAE should have similar shapes when measured consistently. In this chapter, a novel psychophysical method suitable for measuring both types of effects using identical procedures is introduced. The experimental materials, methods and results for the TAE are first described in detail in order to see if this method is compatible with the previous experimental results. Then, the methods and results used to apply this method to the FAE are shown. Finally, similar results for both the TAE and the FAE are shown.

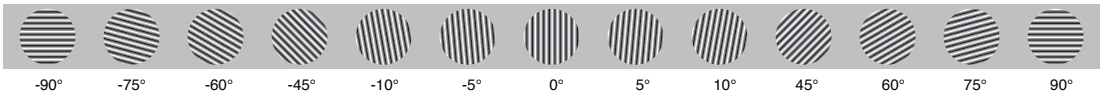


Figure 5.1: Example of grating stimuli for the TAE experiments, covering all orientations.

5.1.1 Participants and apparatus

Two adult male participants were tested for the TAE: the experimenters, CZ and JB. CZ is a 28-year old Chinese male and has been living in a Caucasian face environment for more than a year. JB is a Caucasian male and has had a lifetime of experience with Caucasian faces, such as those used in the experimental stimuli.

The stimuli were presented on a 17-inch 85Hz CRT monitor viewed at roughly $16^\circ \times 19^\circ$ from roughly 45 centimetres.

5.1.2 Stimuli

The stimuli used in the TAE experiments were sine gratings with orientations ranging from -90° to $+90^\circ$, at an approximate spatial frequency of 0.6 cycles per degree (see **Figure 5.1** for an example).

5.1.3 Procedure

A common method for measuring the TAE in previous experiments was to ask the participant to adjust a reference line at a distant location until it appeared parallel to a test line at the adapted location (Mitchell and Muir, 1976). Procedures of this type are not practical for measuring the FAE, because the FAE shows significant transfer across retinotopic locations (Afraz and Cavanagh, 2008). In order to test both FAE and TAE under the same paradigm, a new approach was devised. TAE and FAE were measured by asking participants to make a two-alternative forced choice (2AFC) along a natural perceptual boundary — between left or right vertical grating (TAE) or between male and female faces (FAE). Aftereffects were defined as shifts in this boundary after adaptation. This way, TAE and FAE can be measured using a single paradigm.

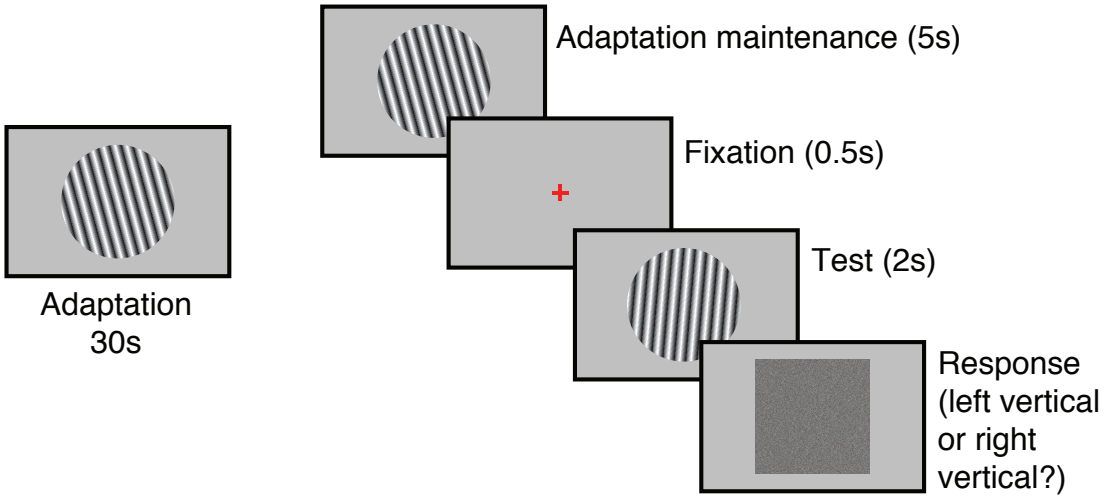


Figure 5.2: Two-alternative forced-choice (2AFC) paradigm for TAE experiment (same as the FAE paradigm except for the stimuli and the response criterion). Participants first became adapted to a grating stimulus and then were tested on random ambiguous test gratings near their perceptual boundary (vertical) so that an updated perceptual boundary could be estimated using a psychometric function on their responses.

The adaptation and test stimuli used in each TAE experiment were drawn from the range of possible grating continuums discussed above (**Figure 5.1**). Each test stimulus consisted of a randomly selected test grating chosen from near the perceptual boundary (vertical) so that the test would be sensitive to changes in the boundary. This procedure is illustrated in **Figure 5.2**.

For each participant, the existing perceptual boundary (vertical grating) was first measured in a baseline block of trials, and then a series of adaptation blocks measured the effects of adapting to different stimuli. Adaptation blocks were the same as baseline blocks except they had extra adaptation and maintenance periods.

In the baseline blocks, test stimuli were chosen uniformly from a range of -4 to +4 degrees around a true vertical grating (orientation 0). Once the unadapted perceptual boundary (vertical) was determined in the baseline block (see details in the next section "Data Analysis"), test stimuli for the adaptation blocks were chosen from a range of -4 to +4 degrees around the baseline boundary for that participant. The stimulus

ranges used in the baseline and adaptation blocks were always identical, as the category boundary of vertical orientation between participants was very stable and almost identical.

Apart from the narrow range around the boundary, two additional stimuli were added to each adaptation test block, significantly different from the perceptual boundary (vertical grating). These two stimuli were -45 and +45 degree orientation gratings. They were added to provide a brief respite for the participants during these difficult experiments, and were clearly recognisable as left or right of vertical.

At the start of an adaptation block, the adaptation stimulus was shown for 30 seconds (adaptation period). Then for each trial, adaptation was topped up for 5 seconds (maintenance period), followed by a 0.5-second fixation mark, a 2-second test stimulus, and then a noise pattern as a signal that a response was required. Once the noise pattern appeared, the participant indicated whether the test grating was left or right vertical (**Figure 5.2**).

Each adaptation block lasted between 10 and 20 minutes, depending on the response time of the participant. Blocks were limited to one per day, to reduce the transfer of adaptation across blocks.

5.1.4 Data analysis

The data from all the trials in each block was collected and used to fit a sigmoidal psychometric function from which the current perceptual boundary could be estimated. The aftereffect for a given stimulus value was the difference between the adapted perceptual boundary in this block and that of the baseline, leading to one data point in a TAE curve for a given orientation (**Figure 5.4a**).

Each data point in **Figure 5.4a** represents the results for one block of trials. Over the course of a block, each stimulus was presented multiple times in order to measure the reliability of the response. In a TAE block, 14 test points (12 for baseline) were typically used, and each point was repeated 8 times. Thus, in total 112 (96 for baseline)

trials were conducted in a TAE block. All the trials were conducted in a random order within a block. Apart from the two extra test stimuli for adapted blocks, the other aspects of the test stimuli were the same for the baseline and adapted blocks.

For each trial in an experimental block, the participant was asked to judge if the grating stimulus was left or right of vertical. The number of times "right" was selected was counted for each stimulus. After all eight possible responses had been collected for a block, the response rates r_i were calculated for each stimulus i in a TAE block:

$$r_i = \frac{N_i}{8} \tag{5.1}$$

where N_i stands for the number of times *right* was selected for stimulus i. Then a sigmoidal psychometric function model f was fit to the response rate across all test stimuli:

$$f(s) = 0.5 + \frac{1}{\sqrt{\pi}} \int_0^{\frac{s-t}{\sqrt{2}b}} e^{-x^2} dx \tag{5.2}$$

where s is the stimulus test range (orientations -90° to $+90^\circ$), and t and b are the threshold and slope parameters of the psychometric curve. The measured perceptual boundary is the x-axis value (threshold) where the psychometric curve has a value of 0.5 on the y-axis, representing perception of a vertical grating. An example of the raw data and the fitted psychometric curve for a TAE block is shown in **Figure 5.3**. This is an adaptation block (participant CZ adapted to 5°), and the measured new boundary is 2.79°.

This psychometric curve model was fit by minimising the residual sum of squares (RSS) between the collected rates r_i and model value f_i for each stimulus i:

$$RSS = \sum_{i=1}^{n} (f_i - r_i)^2 \tag{5.3}$$

where n denotes the total number of test stimuli used in a block. The residual standard deviation was used to quantify the quality of the fitting:

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-p}} \tag{5.4}$$

where $p = 2$, denoting the number of parameters in the model.

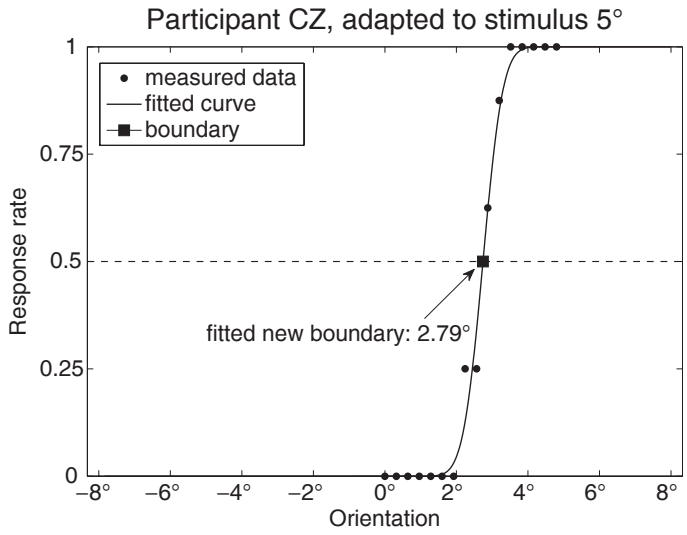


Figure 5.3: Fitted psychometric curve for one of CZ's adapted blocks (adapted to 5°). The value of this psychometric curve at a y-axis value of 0.5 represents the threshold between male and female perceptual judgments in this condition. The difference between this threshold and the value measured for the baseline represents the amount of aftereffect for this condition (adapted to 5°).

5.1.5 Results

The TAE results for two participants are shown in **Figure 5.4a**. These results are similar to those from classical TAE experiments (Mitchell and Muir, 1976), which suggest that this paradigm is comparable to earlier methods, while allowing testing with any type of stimulus.

To evaluate how consistently the participants were able to perform the task, so that problems like fatigue would be evident, the average slopes of the fitted psychometric curves for both participants were measured. The slope of a psychometric curve is a measure of the participant's discriminability, i.e., ability to judge between two categories of stimuli. **Figure 5.4b** shows that the discriminability for each participant has a small standard error of measurement (S.E.M.; plotted as error bars), indicating that it stayed largely constant over the course of the experiments, and thus was not seriously affected by fatigue or similar issues.

The newly devised experimental paradigm is thus compatible with previous TAE

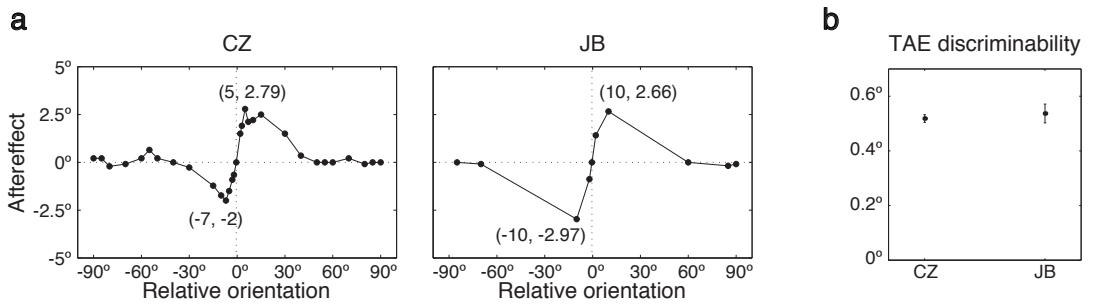


Figure 5.4: Results of TAE experiments. **(a)** Aftereffects measured as shifts in the perceptual boundary (vertical) after adaptation. TAE results for the two participants tested are similar to previously published TAE results (e.g., Mitchell and Muir, 1976; **Figure 4.4d**), verifying that the new paradigm tests similar mechanisms. The average residual standard deviation $\hat{\sigma}$ for all the data points in the two plots is 0.0317. **(b)** Participants' discriminability (slopes of the psychometric curves) for the TAE stimuli. Data points show each participant's average discriminability, with the standard error as error bars. Like the result for the FAE experiment (**Figure 5.7c**), these show that the discriminability for the two participants was largely constant in the experiments, and was not seriously affected by fatigue or similar issues.

experiments such as those conducted by Mitchell and Muir (1976), and therefore it is indeed testing adaptation aftereffects. In the next section, the same method will be applied to the FAE in order to see what kind of aftereffect curve can be obtained and whether there is any similarity between the high-level and low-level aftereffects.

5.2 Face gender aftereffects experiment

All the methods for the FAE experiments were the same as for the TAE; they only differed in stimuli (faces instead of oriented gratings) and response criterion (male or female instead of left or right of vertical). The natural boundary of an androgynous face (i.e., the boundary between a male and female face), fitting the same role as the boundary of a vertical grating in the case of the TAE experiment.

5.2.1 Participants and apparatus

Five adult male participants — the experimenters, CZ and JB, plus naïve participants, JA, CB, LW and ZK — were tested for the FAE. Their ages range from 23 to 39 years old. JA, CB and JB are Caucasian and have had a lifetime of experience with Caucasian faces such as those used in the experimental stimuli. CZ, ZK, and LW are Chinese. CZ and ZK had been living in a Caucasian face environment for more than a year, as LW had been living in a Caucasian face environment for four months.

As in the TAE experiment mentioned in the previous section, the stimuli used in all the experiments were presented on a 17-inch 85Hz CRT monitor viewed at roughly $16^\circ \times 19^\circ$ from roughly 45 centimetres.

5.2.2 Stimuli

The set of face stimuli used in this FAE experiment was the same as in the face gender modelling work described in Section 4.2.1. There were 501 composite faces generated using the methods described by Tiddeman et al. (2001), from 150 young adult male and 150 young adult female Caucasian photographic faces collected by Ian Penton-Voak at Stirling University (Penton-Voak et al., 2006). This face continuum covered the morphing strengths from -2 and +2, generating a wider variety of face shapes than the -1 to +1 range used in previous studies utilising this face set. A sample of the generated synthesised face continuum is shown in the top row (labelled "original") in **Figure 5.5**. It can be seen that even the most extreme faces used here (e.g., -2 and +2) have very strong masculine or feminine features, but are still recognisable as human faces. The face stimuli used a darker grey background than the gratings, to provide high-contrast edges that make the faces salient. More technical details about the face generation method can be seen in Section 4.2.1.

As controls, three additional different versions of the face stimuli were used for different participants, but all of the results are discussed together later in the results

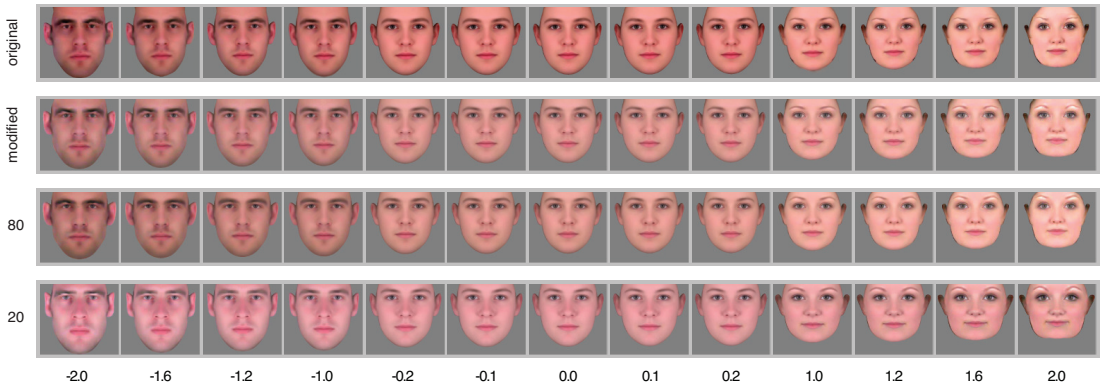


Figure 5.5: Example face stimuli used in each FAE experiment, covering a wide range of male and female faces. Each row shows 13 examples out of the 501 faces making up each continuum. The "original" face continuum (top row) was used for participants CZ and JA. The "modified" continuum (see text) was used for CB. The row marked "80" was generated like the modified continuum, but using only 80% of the faces. The remaining 20% were used for the continuum in the last row. Participants LW and ZK were trained on the 20 continuum and tested on the 80 continuum.

section because no qualitative or quantitative differences in results were found between the participants or stimulus sets. The first version of the face morph continuum that was created is shown in **Figure 5.5** "original", used for participants CZ and JA. To see if the results obtained from the "original" face set were affected by the slight left-right asymmetry in face shape and the overall differences in skin tone between male and female images, the "modified" face set in **Figure 5.5** was generated. For the "modified" face set, the key locations for the face morphing were averaged across the vertical midline, enforcing symmetry, and the overall trend of colour changing with morphing strength was reduced by manually adjusting the colour balance of the two male and female average images to match that of the population average more closely. As mentioned, the results from the participants using the "modified" face set were similar to the "original" face set.

For both the original and modified datasets, participants were tested with faces drawn from the same morphing continuum as the adaptation stimuli. With faces gen-

erated along a smooth continuum, many of the low-level features will vary systematically along with gender, such as the shape of the eyes, face outline, and mouth, the brightness of skin textures, etc. To test whether low-level adaptation to the individual features could explain the observed FAE, the original face photographs were split into one set of 80% (120 males and 120 females) and another set with the remaining 20% (30s male and 30 females). A face continuum was then generated for each of those separate datasets by averaging and morphing as described above; the generated continuums are shown in **Figure 5.5**, rows "80" and "20". Participants LW and ZK were then adapted to faces from the 20% set, but tested on faces from the 80% set, so that the low-level features would only matter if they were reliably indicative of facial variances across individuals.

In addition to these three different face continuum datasets, a fourth condition was used to investigate contributions from low-level effects, by varying the size between adaptation and test phases. The stimuli in the fourth FAE experiment were the same as the "modified" face set in **Figure 5.5**, except the test faces were half the width and height of the adaptation face. I.e., the overall size of the test face was 25% of the area covered by the adaptation face, disrupting most of the low-level features but keeping the face appearance and identity intact. The test and adaptation faces were aligned along the centre point between two eyes, where the fixation point was placed. Otherwise, the experimental procedure was identical to the other conditions, as described below.

5.2.3 Procedure

The same paradigm and procedure as for TAE in Section 5.1.3 were used in the FAE experiments. The details are described here and focus on the differences with the TAE experiment.

The purpose of the FAE experiment is to measure the aftereffects as shifts in the boundary of an androgynous face after adaptation. To achieve this, the participants

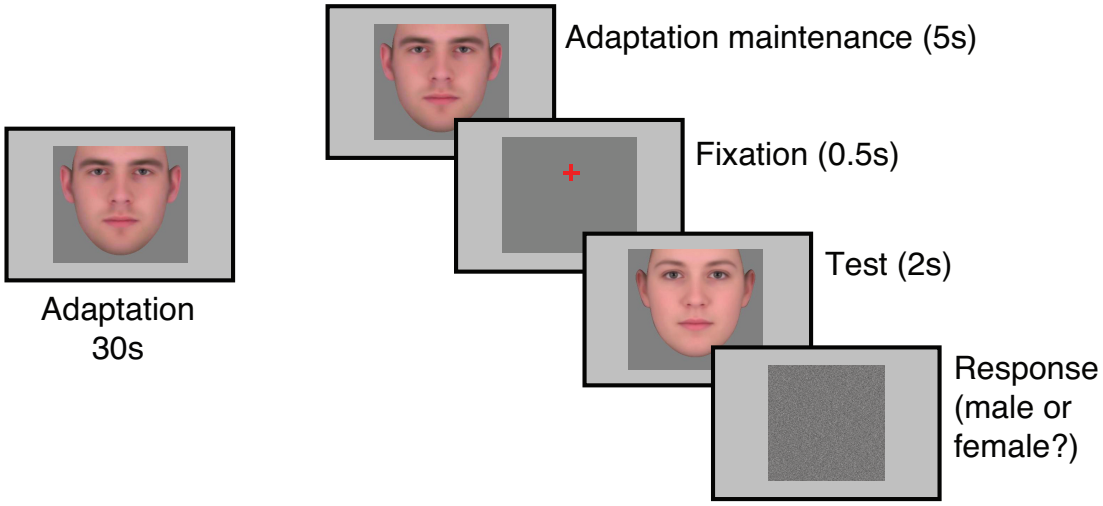


Figure 5.6: Two-alternative forced-choice (2AFC) paradigm for the FAE experiment. Participants were first adapted to a face stimulus, then were tested on random ambiguous test faces near their perceptual boundary (androgynous) so that an updated perceptual boundary could be estimated using a psychometric function on their responses.

were asked to make a 2AFC choice between faces that were very close to an androgynous face. The adaptation and test stimuli were drawn from the range of possible face continuums discussed above (**Figure 5.5**). Each test face consisted of a randomly selected ambiguous test pattern chosen from near the androgynous face so that the test would be sensitive to changes in the boundary.

Like in the TAE experiments, the participants also conducted baseline and adaptation blocks. In the baseline blocks, test stimuli were chosen uniformly from a range of -0.24 to 0.24 morphing strengths around the perceptual boundary (androgynous face at morphing strength 0). Once the unadapted perceptual boundary was determined, test stimuli for the adaptation blocks were chosen from a range of -0.48 to 0.48 morphing strengths around the baseline boundary for that participant. In the FAE experiments, each participant had a different baseline ("androgynous") category boundary initially. To best suit the specific baseline for that participant, an appropriate test range was chosen individually — some were tested in the range of -0.08 to 0.48 and others from -0.48 to 0.08. Of the six FAE participants, ZK, LW and JB used the same stimulus

range for baseline and adaptation, while the three others used different ranges. Note that in some cases, it has previously been shown that changing the range in this way could induce a shift in the point of subjective equality (PSE), i.e., the gender boundary used for testing (see Poulton, 1974 and Laming, 1997). Such range effects could potentially bias the boundary points leftwards or rightwards (see crossing data points in **Figure 5.7a**) in those participants, but they would not be expected to alter the overall shapes of the aftereffect curves, which are the focus of this study.

The two additional stimuli added to each adaptation test block were -1.2 and 1.2 morphing strengths. In an adaptation block, the adaptation period and the maintenance period were at the same length as the TAE. As for the TAE, participants were allowed to view the adaptation face freely during the adaptation period, and were instructed to move their eyes around to avoid afterimages and other distracting low-level effects. In each trial, once the noise pattern appeared, the participant indicated whether the test grating was male or female (**Figure 5.6**). Before a block, the participants were instructed to judge the face holistically and to use only their first impression.

An FAE adaptation block was typically 2 to 5 minutes longer than a TAE adaptation block, because participants appeared to need more time to make a high-level judgment of face gender. As for the TAE experiment, blocks were limited to one per day to reduce the transfer of adaptation across blocks. This procedure is illustrated in **Figure 5.6**. In general, apart from the stimuli and the response criterion, the 2AFC paradigm and experimental procedure were identical to those used in the TAE described in Section 5.1.3.

5.2.4 Data analysis

The way of analysing the experimental data was identical to the methods for TAE illustrated in Section 5.1.4. As for the TAE, the procedure was to collect data points from all the trials, then fit a sigmoidal psychometric function and thus estimate the current perceptual boundary. Then, the aftereffect for a given stimulus value was the differ-

ence between the adapted perceptual boundary in this block and that of the baseline, leading to one data point in an FAE curve for a given adaptation morphing strength (**Figure 5.7a**).

Over the course of an FAE block, each stimulus was presented 8 times in order to measure the reliability of the response. 18 test points were used for an adaptation block and 16 for a baseline block. As mentioned above, in the experiments, each participant's test ranges were slightly different, so at most there could be 20 or at least 17 test points in the adaptation blocks. Therefore, in total 136 to 160 (128 for baseline) trials were conducted in an FAE block.

The technical details for fitting a sigmoidal psychometric function and determining the perceptual boundary were the same as those described in Section 5.1.4.

5.2.5 Results

As mentioned above, the specific face stimuli for each participant differed, but the FAE results are presented as one group because the FAE for every participant matched the TAE-like S-shaped curve predicted from the model described in the previous chapter (see experimental results in **Figure 5.7a** and model predictions in **Figure 4.5**). The effects were similar between participants CZ and JA using the "original" face set and participant CB using the "modified" face set. Similar curves were also found for ZK and LW, indicating that the results hold even for perceptually different adaptation stimuli (compare the -2.0 and +2.0 faces for the 80% and 20% datasets). These results presumably reflect the fact that the average faces (morphing strength 0) are similar for all of the datasets, as would be expected for any averages of large numbers of faces drawn from the same distribution. Finally, even when the stimulus size differed by a factor of four in the area between the adaptation and the test (participant JB), a similar S-shaped curve was observed, though the aftereffect strength was lower in this case.

Figure 5.7b shows that the presence and location of the peak and valley in the S-shaped curve was highly consistent across individuals, with aftereffects significantly

stronger at morphing strength ± 1.2 than at ± 2 (one-tailed paired Student t-test; $p < 0.0126$). Put together, all these results indicate that FAE magnitude decreases after reaching a maximum value in either direction of difference from the mean face. The control experiments also demonstrated that this aftereffect curve is qualitatively invariant to identity and size difference, as would be expected for high-level effects.

As shown in **Figure 5.7c**, the average slopes of the fitted psychometric curves for 6 participants both have a small standard error of measurement, indicating that their discriminabilities stayed constant over the course of the experiments, and thus that they were not seriously affected by fatigue or similar issues. This result is consistent with the TAE shown in **Figure 5.4b**, indicating that both paradigms were stable. These results are consistent with previous reports that adaptation to extreme face conditions does not facilitate discrimination around the average face (Rhodes et al., 2007).

More importantly, it can be seen that the FAE results (**Figure 5.7a**) are strikingly similar to the TAE results (**Figure 5.4a**) in the previous section — with the adaptation stimuli being further away from vertical grating or being more masculine or feminine, aftereffects, computed as shifted perceptual boundary, steadily decrease after reaching peak value around the middle adaptation orientation or face morph. As they were based on the same paradigm, and only differed in the stimuli and response criterion, these results suggest that the two kinds of aftereffects may share similar neural mechanisms. This result is consistent with existing psychophysical studies of the human face (Webster et al., 2004; Little et al., 2005; Leopold et al., 2001; Jeffery et al., 2010) that found aftereffects that monotonically increase in magnitude with difference away from an average face, because those studies used a smaller range of faces and thus only tested the rising portion of the curve.

These results also have a similar shape to the model predictions from Sections 4.2 and 4.3. This could suggest that lateral inhibitory Hebbian learning (the LISSOM model in Section 4.2) or reduced responsiveness of the face's most activated neurons (the exemplar-based model in Section 4.3) could be the main contributors to face adap-

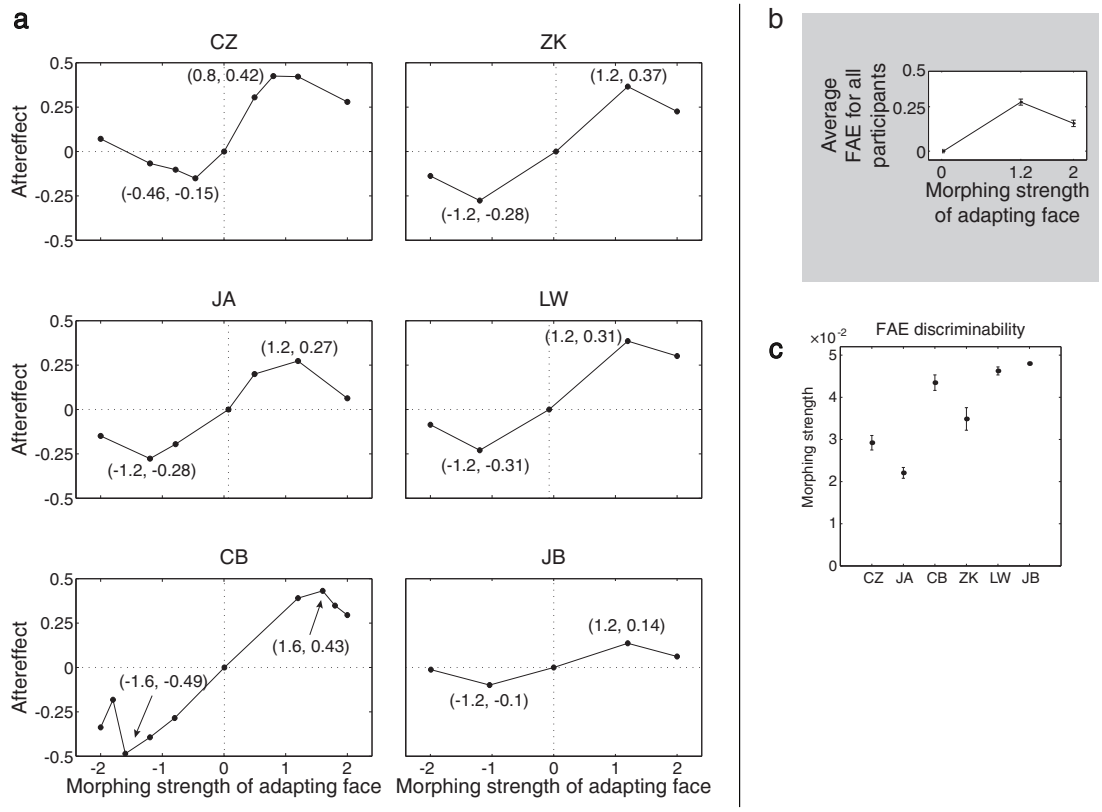


Figure 5.7: Results of FAE experiments. **(a)** Aftereffects measured as shifts in the perceptual boundary after adaptation. FAE results for the six participants tested suggested broad tuning, as expected. In all cases the FAE magnitude at extreme morphing values decreased after reaching a maximum as for the TAE (Figure 5.4a), as predicted from the models but contrary to previous reports. The average residual standard deviation $\hat{\sigma}$ for all the data points in the six plots is 0.0671. **(b)** The magnitude of the FAE averaged across all six participants and across both male and female test faces (for a total of 12 measurements), with error bars indicating the standard error of measurement. The value at morphing strength 2.0 is significantly lower than the value at 1.2 (one-tailed paired Student t-test; $p < 0.0126$), indicating that the S shape of the aftereffect was reliable across participants. The horizontal error bar at aftereffect zero shows the small variation between participants in the perceptual boundary at the baseline. **(c)** Participants' discriminability (slopes of the psychometric curves) for the FAE stimuli. Data points show each participant's average discriminability, with the standard error as error bars. These results show that the discriminability for each participant was largely constant in the experiments, and was not seriously affected by fatigue or similar issues, as for the TAE experiment (Figure 5.4c).

tation. These models used the same approaches to account for the TAE, while the experimental TAE and FAE curves presented in this chapter showed similar shapes. Together, these results provide support for the idea that high-level aftereffects like the FAE may be processed in a similar way to low-level aftereffects like the TAE.

5.3 Discussion

The above TAE results showed that the TAE experiment gave results that were similar to previous low-level aftereffects experiments. The above FAE results indicated that the FAE curve is not ever-increasing in opposite directions from the average face; instead, it reliably decreases once the adaptation faces are sufficiently dissimilar to the average face. As described in Section 5.2.2, much effort has been made to avoid interference of low-level factors, suggesting that the FAE experiments tested high-level aftereffects. The similarity of FAE and TAE shapes indicate that low-level and high-level adaptation aftereffects could be based on similar neural mechanisms. Such a similarity between low-level tilt aftereffects and high-level face gender aftereffects had not been demonstrated previously.

As previously mentioned, the stimulus faces were synthesised using a linear PCA method. Of course, each resulting face continuum may or may not be perceptually linear, i.e., with similar differences in face appearance at each pair of neighbouring points on the continuum. Even so, as long as the continuum is perceptually monotonic (i.e., with faces consistently becoming more female as the morphing strength increases), finding an S-shaped curve will not be affected by such nonlinearities. Nonlinearities in a monotonic dimension can change the slope of specific parts of the S-shaped aftereffect curve, but could not flatten it as a whole or change the number or sign of the peaks. The required monotonicity can be verified visually for each dataset in **Figure 5.5. All observers so far have agreed that the continuums are monotonic.**

As mentioned in the experimental procedure above, each experimental block typ-

ically took 15 to 20 minutes for the TAE and 15 to 25 minutes for the FAE. The time difference was due to the larger number of trials and the slower average reaction times for the FAE stimuli. In pilot trials, the chosen duration was found to provide a good balance between getting sufficient data to measure the boundary precisely for that block, while avoiding participant fatigue that could compromise the quality of the results. As each participant was asked to conduct 5 to 10 blocks in total to get enough data points for the curves in **Figure 5.7**, each FAE curve represents 100 to 250 minutes on average spent performing this difficult task over 5 to 10 days. Thus, it is not practical to repeat the entire task enough times to be able to estimate the statistical significance or error bars for each participant using standard methods. Over the course of the 10 or so repetitions of the entire task (17 to 33 hours over 50 to 100 days) that would be required for such tests, participants would be likely to adapt to many aspects of the tests other than those being measured explicitly. Thus, it would be difficult to interpret the results, even if time considerations and personal health issues did not preclude running such tests. An alternative, less time-consuming method to measure these effects is in progress to use Bayesian methods to limit the range of test stimuli and approach the participant's boundary more rapidly (Watt and Andrews, 1981).

The experimental results presented in this chapter also link the models described in Chapter 4. Two kinds of models, LISSOM-based and exemplar-based multichannel, can thus both account for the TAE and the FAE, as the model predictions are both consistent with the experimental ones in this chapter. The consistent result for the TAE and the FAE with the LISSOM-based networks shows that the lateral inhibitory plasticity can account for neural adaptation for both low- and high-level effects. Both LISSOM and the exemplar-based multichannel model, as will be discussed in the final chapter, provide alternative theories to account for face aftereffects, both unlike the "norm-based" encoding theory, and the evidence in this chapter should prompt a re-examination of previous claims for norm-based encoding.

Chapter 6

Modelling face emotion aftereffects

This chapter first briefly reviews previous experimental work on the transfer of face emotion aftereffect from low- to high-level cortical areas. Then the details of the modelling work duplicating this aftereffect are introduced, including the model architecture, face generator, decoding methods and the aftereffects simulation. The limitations of the current model are also discussed, followed by the results of the modelling.

6.1 Introduction

The experimental work described in the previous chapter shows that the high-level face gender aftereffect may be a result of similar neural mechanisms as in the low-level tilt aftereffect. Much effort was made to reduce the possible interference of low-level features (see Sections 5.2.2 and 5.2.3 for more details about the control experiments on face brightness, identity and size), without eliminating the effect, which gives us confidence that the FAE experiments mentioned in the previous chapter did indeed test high-level aftereffects.

Even so, it is clear that responses to high-level stimuli like faces will involve both low-level and high-level cortical processing (see Section 2.1 for a review). When a visual stimulus is presented in the retina, the visual information might pass through low-level subcortical and cortical areas (retina, LGN, V1, etc.) to high-level areas (IT,

FFA, etc.). Thus it is interesting to ask a further question — what effect will low-level adaptation have on high-level adaptation aftereffects? It is reasonable to assume that the processing of high-level stimuli will include effects of low-level adaptation. For example, although V1 may suffice for representing and encoding an oriented grating stimulus, a range of visual areas from V1 to FFA are expected to be involved in representing and encoding a face stimulus.

As a first step, it is important to determine if a fraction of the observed gender FAE is due to low-level rather than high-level adaptation. As reviewed in Section 2.2, Xu et al. (2008) provided experimental evidence of such a transfer across levels. In brief, they showed that adaptation to curved lines can affect high-level perception of emotional in faces, presumably by modifying the perception of mouth curvature. Their result shows that aftereffects can propagate up the cortical hierarchy. For the patterns they had tested so far, the effect increases as the adaptation pattern curvature increases, just as had been found for high-level FAEs in previous studies.

Given the previous results in this thesis, one may ask whether such a transfer of emotion aftereffects should also lead to an S-shaped curve as in the TAE and FAE. In the experiments conducted by Xu et al. (2008), the ranges of the curvatures of curved lines, cartoon faces and real faces are all limited, and therefore it may be possible to find an S-shaped aftereffect if the full range is tested. This chapter will show cartoon face modelling results suggesting that an S-shaped curve is indeed predicted by the model, which can be tested in future experimental work.

6.2 Methods

In this section, the details of the modelling transfer of face emotion aftereffects using LISSOM are described. The model architecture, stimulus generator, decoding method and simulated adaptation are shown, followed by the modelling results and discussion.

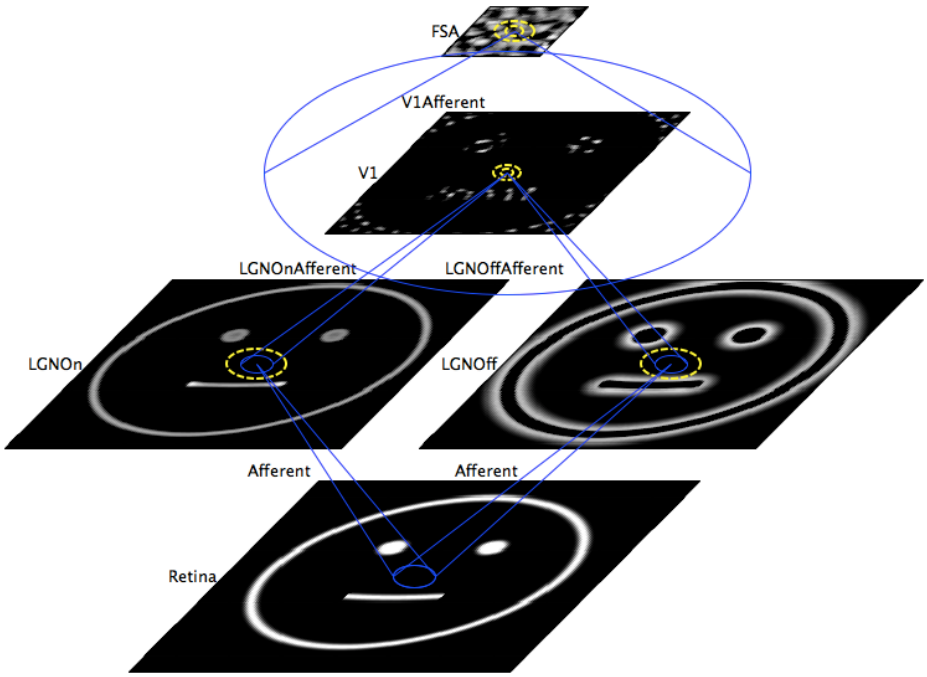


Figure 6.1: Hierarchical LISSOM architecture for the simulation of cartoon-based face emotion aftereffects. This model extends the simple FSA model described in Sections 3.2.1 and 4.2.1 by adding a V1 sheet. A neuron in each upper sheet received afferent input from neurons in the lower sheet, and neurons in the same sheet are laterally connected to each other. **Top**: the face-selective area (FSA) sheet represents face space; **second row**: V1 sheet simulates V1; **third row**: LGNOn and LGNOff sheets simulate RGC/LGN; **bottom**: photoreceptor input sheet.

6.2.1 Multi-layered model architecture

As described in Chapters 3 and 4, LISSOM was successful at duplicating and predicting the face identity and gender aftereffects, and so the same architecture was also used to test emotion aftereffects (see Figure 6.1).

Apart from the sheets of photoreceptor, RGC/LGNs and FSA (face-selective area) as in Section 4.2.1, a V1 sheet was added before the FSA. This multi-level model was then used to simulate the cross-cortical transfer of aftereffects.

The modelling of this transfer of emotion aftereffects assumes that the high-level face emotion aftereffects found in experiments were due to changes in the representation of lines in V1. After adapting to the curved lines, V1 responses to subsequent

input stimuli differ due to changes in the responsiveness of V1 neurons. Thus, the same input stimuli before and after adaptation will lead to slightly different V1 activation patterns. The curved lines used in adaptation are identical in shape to the mouth in a cartoon face, so nearly the same set of V1 neurons that respond to the curved lines will respond to the cartoon face. As the adaptation duration was very short, high-level neural wiring and the face preference map are assumed to retain their pre-adaptation responsiveness. In this way, the different V1 outputs lead to changes in activation patterns on the FSA sheet. Such misrepresentation is manifested as the cross-cortical transfer of emotion aftereffects in this model.

The afferent connections from RGC/LGN to V1 sheet and the lateral connections in the V1 sheet were exactly the same as the standard V1 LISSOM model described in Miikkulainen et al., 2005 Section 4.2. The parameters for the afferent connection size and learning rate from V1 to FSA sheet, and the lateral connection sizes and learning rates in the FSA sheet were slightly different from the previous model described in this thesis. This difference was mainly because that in the previous models, FSA was used to train photographic faces; while in this model, FSA was used to train schematic cartoon face patterns (see the samples of cartoon faces in Figure 6.2). These parameters were best tuned for this set of stimuli. These changes allowed the FSA neurons to process the full eyes and mouth region at the cartoon face.

6.2.2 Cartoon face generator

Both the curved lines (used in adaptation) and cartoon faces were rendered by the Topographica simulation running the LISSOM network. The Xu et al. (2008) paper did not provide complete numeric parameters of the curved lines and cartoon faces, but the shape of the curved lines and cartoon faces were designed to approach the images shown in the paper as closely as possible. Examples of the curved lines and cartoon faces are shown in Figure 6.2.

The faces and curved lines on the left-hand side of Figure 6.2 present sad emotions,

Figure 6.2: Examples of the mouth curve and cartoon face stimuli used in the model. The mouth curves in **a** are the same as their corresponding mouth portion in the faces of **b**. Mouth and face shapes use the same curvature scale, in units of $1/deg$. Negative values mean a sad (concave) emotion and positive mean a happy (convex) emotion. The line length is kept constant throughout.

and those on the right side happy emotion. Positive and negative values were added in order to show the differences clearly, even though in the Xu et al. (2008) paper, sadness and happiness were represented by the left-to-right order only. Thus, in this chapter, a negative happiness value equals left-side emotions, and a positive happiness value equals right-side emotions. Black lines on the plots indicate high photoreceptor activations (the black lines on white background are only for the purpose of visualisation in this paper, as shown in Figure 6.2, Figure 6.4 and Figure 6.8b). Unlike the paper, no real face stimuli are shown here, which will be discussed in Section 6.4.

Following the convention of how the curved lines and cartoon faces were constructed in the Xu et al. (2008) paper, this chapter defined the curvature of both the curved line and face mouth (in the cartoon face) as $1/deg$. As illustrated in Figure 6.3, curvature is defined by

$$1/deg = \frac{2l}{\pi r}, \tag{6.1}$$

where l denotes the length of the curved line or face mouth, and r denotes the radius of this fraction of curve. During the simulation, the curve length l was kept intact, in order to ensure that the number of neurons involved in the adaptation was roughly the same for all the adaptation conditions. Thus, the curvature $1/deg$ increases (in both

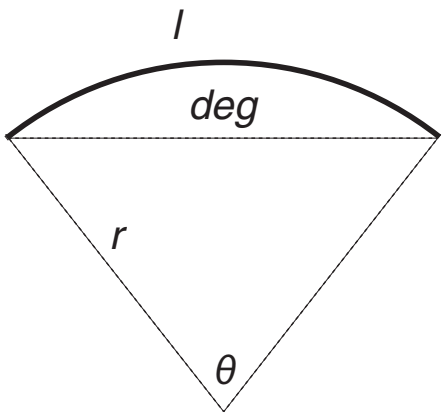


Figure 6.3: Diagram of the parameters of a mouth stimulus. l stands for the mouth length, deg for the mouth width ($1/deg$ as the curvature unit), r for the radius of this arc, and θ for the degree of this arc.

negative for sadness and positive for happiness) while the radius r is becoming shorter. For all the curved lines and cartoon faces, the position (x, y) was also kept intact given a curvature $1/deg$, where x stands for the horizontal centre of the curve or face mouth, and y for the vertical centre of the height of the curve or face mouth. This way, all the curved lines and face mouths were in the same region for all the test cases and adaptation conditions. Keeping the stimulus well-registered helps to reduce irrelevant variation in the responses caused by activating different specific subsets of neurons.

6.2.3 Decoding and aftereffects simulation

As mentioned earlier, in this multi-layered network, the photoreceptor input was first filtered by the V1 sheet, and then the V1 output was fed into the FSA sheet, and finally the activity in the FSA sheet was read and decoded as the "happiness" of the stimuli. Examples of the cartoon face stimuli and their corresponding FSA responses are shown in Figure 6.4.

Decoding the perceived happiness was done using the correlation comparison decoding method from Leopold et al. (2001), as described in Section 3.2.3, because the purpose of this chapter is to duplicate the result in the Xu et al. (2008) paper. Xu et al.

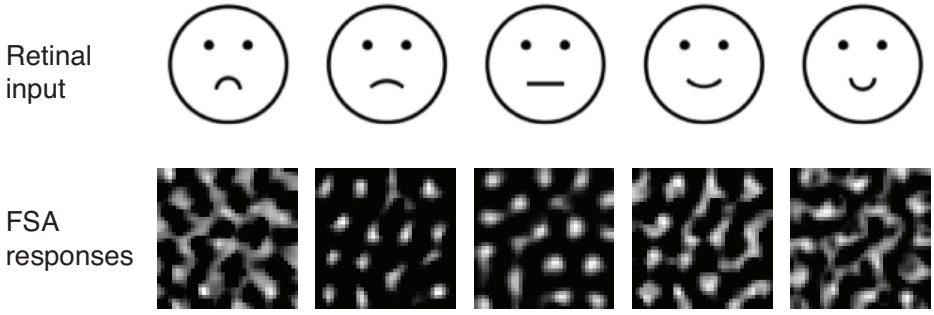


Figure 6.4: Examples of the cartoon face stimuli (as photoreceptor input to the model) and their corresponding FSA responses. **Top**: examples of cartoon faces from the saddest (curvature -1) to the happiest (curvature 1); **bottom**: corresponding FSA response elicited by the faces after the V1 sheet has been trained for 20,000 iterations and then the FSA sheet trained for 3,000 iterations.

(2008) showed the psychometric curves as the main results, not perception shifts for all adaptation conditions, as provided by the indirect local comparison method from Section 3.2.4. The Leopold et al. (2001) experiment also produced psychometric curves that were duplicated in Chapter 3. Therefore, the same correlation comparison decoding method used in Section 3.2.3 was implemented here.

The experimental paradigm from the work of Xu et al. (2008) is shown in Figure 6.5, which is very similar to the paradigm described in Figure 5.6 in Chapter 5. The simulation process described in Section 4.2.2 was able to predict the result of Chapter 5, and so the same process was used here. Put together, in this model, the process of aftereffects simulation is the same as that explained in Section 4.2.2, apart from an extra pre-training process of V1. As mentioned in Section 6.2.1, the aftereffect is produced by the differences in V1 output that lead to further differences in the FSA.

To ease the process of computational simulation, in this model the V1 and FSA sheets are trained sequentially rather than simultaneously. The FSA neurons are also not "aware" of the V1 changes during adaptation, and the FSA neurons are not themselves adapted during stimulus presentation. This approach highlights the contribution of low-level adaptation, because only low-level changes can account for the effects

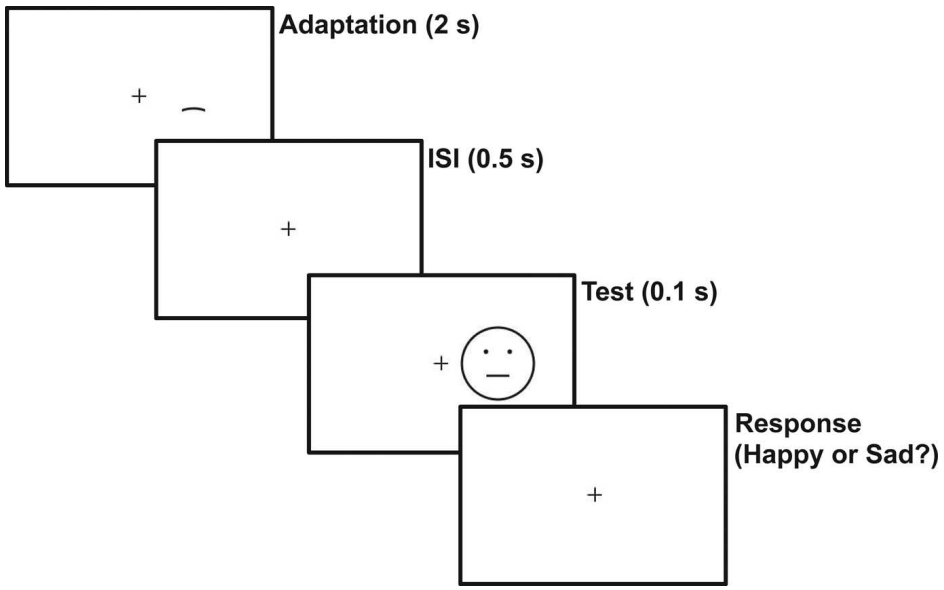


Figure 6.5: 2AFC paradigm of the psychophysical experiments conducted by Xu et al. (2008). This diagram shows the process of adaptation to the mouth curve of test on cartoon face. The procedure of other experiments described in this work is same apart from the different adaptation or test stimuli.

seen. In humans, adaptation would be expected to occur at all levels, but not necessarily at the same rates, leading to "unaware" responses from high-level areas adapting relatively slowly to response-pattern changes below them.

In order to test whether the transfer of aftereffects could arise between curved lines and cartoon faces, four kinds of adaptation conditions were tested by Xu et al. (2008): adaptation to the curved line and test on the cartoon face (denoted c-f), adaptation to the curved line and test on the curved line (denoted c-c), adaptation to the cartoon face and test on the curved line (denoted f-c) and adaptation to the cartoon face and test on the cartoon face (denoted f-f). An additional inverted face condition was also tested in the experiment, in order to see how face inversion would affect the transfer of aftereffects (adaptation to the inverted cartoon face and test on the upright cartoon face, denoted if-f).

The entire training and simulation process, in order, was as follows:

- Pre-adaptation:

1. Train the V1 sheet with natural images;
2. Let V1 respond to a cartoon face (for the c-f, f-f and if-f conditions) or a curved line (for the f-c and c-c conditions);
3. Train the FSA sheet with the V1 output for a random curvature value until the FSA develops an organisation of face space in terms of happiness;
4. For each curvature, produce FSA output from the V1 output;
5. Decode each FSA output from step 4 as an estimate of face happiness forming the baseline;

- Adaptation:

    1. Briefly train V1 with a curved line (for the c-f and c-c conditions), a cartoon face (for the f-c and f-f conditions) or an inverted face (for the if-f condition);

- After adaptation:

    1. Present each cartoon face (for the c-f, f-f and if-f conditions) or a curved line (for the f-c and c-c conditions);
    2. Decode the FSA outputs as estimates of face happiness;

In the pre-adaptation period, V1 was trained with natural images rather than curved lines or cartoon faces, but the FSA was pre-trained with cartoon faces only. This difference is intended to simulate the real roles of V1 and the hypothesised FSA — V1 should process all kinds of visual information (as the natural images contain all types of visual patterns) while FSA is assumed to be dedicated to the face processing.

In the adaptation period, as explained above, only the V1 sheet was involved and the FSA sheet connections were not affected. In particular, only the V1 lateral inhibitory connections were modified, for the same reasons as described in Section 3.2.5.

The post-adaptation period was identical to the pre-adaptation period, except that there were no network training steps.

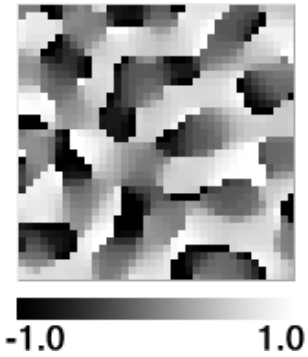


Figure 6.6: Representation of face happiness of the trained model. The happiness preference of each FSA neuron is shown as a 2D sheet, with colour gradient from curvature -1 (black) standing for the saddest (concave) to curvature 1 (white) standing for the happiest (convex). The V1 sheet was first trained for 20,000 iterations, followed by FSA sheet training for 3,000 iterations.

Apart from the above, other elements in the adaptation simulation are identical to the process described in Chapter 3 and 4.

6.3 Results

As in Section 3.2.2, it is necessary to check if the cartoon face generator and the LISSOM network are representing this stimulus type reliably. Accordingly, a face happiness preference map was measured to show the quality of the happiness representation, as illustrated in Figure 6.6.

This map shows each FSA neuron's preferred stimulus value (ranging from -1.0 to 1.0) in the dimension of face happiness. If the model is well-organised by the input stimuli, the self-organising process should have clustered neurons with similar happy or sad preferences together, and thus the model should yield a smooth (in grey-scale colour) preference map and gradually changing receptive fields along a trajectory in the FSA sheet. It can be seen in Figure 6.6 that the map is smooth in grey-scale, and visually there is a relatively uniform distribution for all the preference values. These indicate that this model is well trained and organised for face happiness.

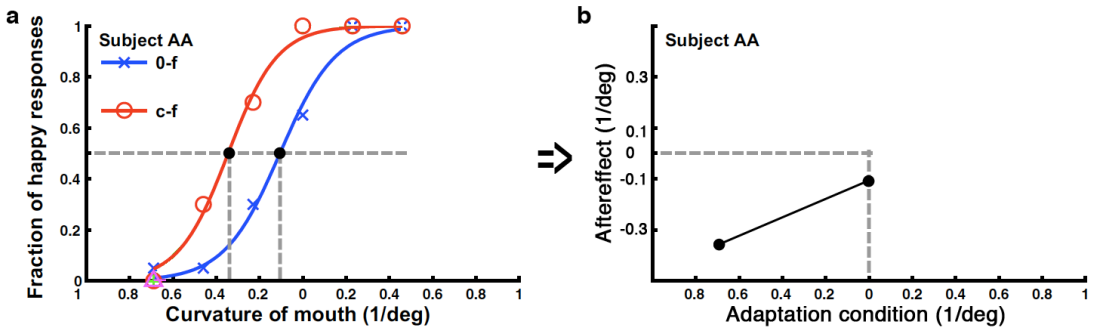


Figure 6.7: Calculating and linking the PSE points from the psychometric curves from Xu et al. (2008). **a**: Psychometric functions from a naive subject for the perception of cartoon faces under the following conditions (adapted from Xu et al., 2008). 0-f, no adaptation (baseline, blue); c-f, adaptation to the most concave curve, identical to the mouth of the saddest face (red). For each condition, the fraction of happy responses was plotted as a function of the mouth curvature of the test faces. The PSE points for these two conditions, i.e., points at the happiness responses fraction 0.5, are shown as solid black dots. **b**: The absolute values of the mouth curvatures are plotted as a function of their adaptation conditions: no adaptation (curvature 0) and adaptation to the most concave curve (curvature -0.69).

Having measured the FSA happiness preference map, it is used to decode the perceived happiness, just as in Section 3.2.3. As already mentioned, the correlation comparison method was used for decoding.

It should be noted that in the Xu et al. (2008) paper, the results were represented by the psychometric curve that represents perceptual sensitivity, but in the model the results are reported as a direct perceptual value. As before, the point of subjective equality (PSE) was used to link them. Figure 6.7 shows that the midpoint of each psychometric curve can be used as aftereffect values, corresponding to the 0.5 perceptual happiness point in the model.

Examples of cartoon faces and curved lines used in the model are shown in Figure 6.8a, and those used in the experiments in Figure 6.8b. Note that the scaling for both cases is normalised so that the saddest and happiest curved lines and cartoon faces correspond to ± 0.46 in the model. In the paper and the model, all adaptation conditions

were either the saddest (-0.46) curved line or cartoon face. The model results for 0-f (baseline), c-f, f-f and if-f conditions are shown in Figure 6.8e, and results for 0-c (baseline), c-c and f-c are shown in Figure 6.8f.

In the results, perceptual boundaries points (PSE) in Figure 6.8c and e with the same colour are correspondent; perceptual boundaries points (PSE) in Figure 6.8d and f with the same colour are also correspondent. It can be seen from the above results that apart from the if-f condition, the PSE points in Figure 6.8e are qualitatively similar to the points in Figure 6.8c, likewise Figure 6.8f and Figure 6.8d — the PSE points for corresponding c-f and f-f conditions shift leftwards to the corresponding 0-f condition, and the PSE points for the corresponding f-c and c-c conditions shift leftwards to the corresponding 0-c condition. For these adaptation conditions, this model replicated the results of the Xu et al. (2008) paper, providing a concrete explanation for how those effects can come about.

In this model, the difference between the adaptation to cartoon face and curve conditions, both with testing on cartoon faces and curves, is not as strong as in the human experiments. The aftereffect for f-f is larger than c-f, but in the case of the model, they are similar, and similarly for c-c and f-c. This difference found in the human experiments may reflect additional adaptation at higher levels that is currently not included in this simple model.

Also, this model did not show strong aftereffects when adapted to an inverted face (the if-f condition). This difference is likely because in the experiment the subject was instructed to judge the face expression and thus likely to make an eye movement towards the mouth; while the model adaptation is local to the upright mouth, the inverted mouth does not activate the same set of retinotopic neurons.

The human experiments by Xu et al. (2008) included several adaptation conditions, but not necessarily a complete picture of the face happiness aftereffect. Does the model predict an S-shaped curved as in Chapters 4 and 5? To test this idea, additional decoded perception curves were computed for the adaptation to “sad” curved lines (Figure 6.9a)

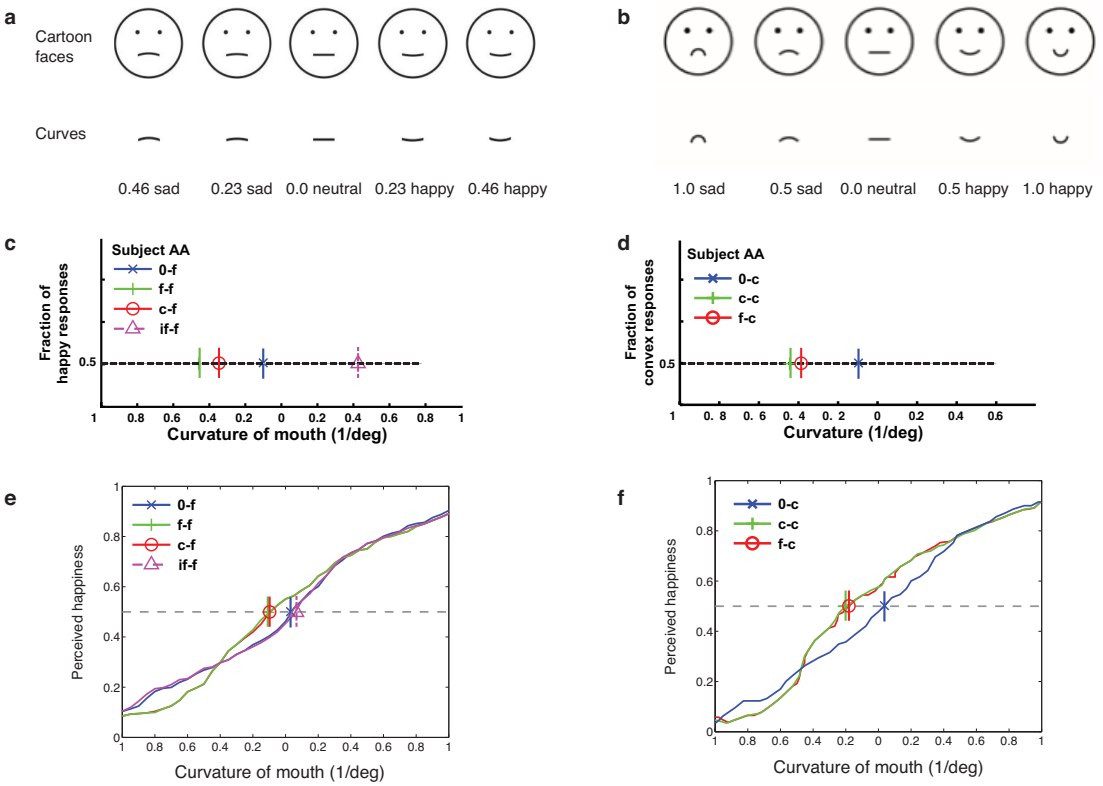


Figure 6.8: Stimuli and results for the emotion aftereffect model, in line with the experimental results for comparison. The figures of (a), (c) and (e) are from the experimental work conducted by Xu et al. (2008). **a**: Cartoon faces and mouth curves used in the experiments. Note that the most extreme sad (curvature -0.69) and happy (curvature 0.69) faces and curves are not shown. **b**: Cartoon faces and mouth curves used in the model. The faces and curves stimuli are in the range between the saddest (curvature -1; semicircle) and the happiest (curvature 1; semicircle). **c**: PSEs, i.e., points at the happiness responses fraction 0.5, extracted from the psychometric functions from a naive subject on the perception of the cartoon faces under the following conditions: no adaptation baseline (0-f, blue), adaptation to the saddest face (f-f, green), to the most concave (saddest) curve (c-f, red), to the inverted saddest face (if-f, magenta dashed). They are plotted along the black dashed line. **d**: PSEs extracted from the psychometric functions from a naive subject on the perception of the mouth curvature under the following conditions: no adaptation baseline (0-c, blue), adaptation to the most concave curve (c-c, green), to the saddest face (f-c, red). Otherwise the same as **c**. **e**: Decoded perception on the test of the cartoon faces under the following conditions. No adaptation baseline (0-f, blue); adaptation to the 0.4 sad cartoon face (f-f, green); adaptation to the 0.4 sad curvature (c-f red); adaptation to the 0.4 sad inverted face (if-f, magenta). For each condition, the decoded happiness was plotted as a function of the mouth curvature of the test faces. The PSE at the happiness 0.5 (neutral emotion) are shown as colour marks (same meaning as in **c**) along the black dashed line. **f**: Decoded perception on the test of the mouth curves under the following conditions. No adaptation baseline (0-f, blue); adaptation to the 0.4 sad cartoon face (f-c, red); adaptation to the 0.4 sad curvature (c-c, green). Otherwise same as **e**.

and to "happy" curved lines (Figure 6.9b). Twenty-one such curves were computed on both sad and happy sides, and their PSE are linked to plot a face happiness aftereffect curve as shown in Figure 6.9c.

It can be seen from the prediction (Figure 6.9c) that the resulting aftereffect curve is also S-shaped, qualitatively like the curve for TAE or face gender aftereffects in Chapters 4 and 5. Note that the 0.4 points on both sides roughly correspond to the saddest and happiest cases in the human experiment. The experimental and modelling paradigms described in this chapter are also similar to those in Chapters 4 and 5 for face gender aftereffect. Thus, to test the model, further human experiments on the most extreme sad and happy conditions (i.e., -1.0 and 1.0) can be performed. If the results are as predicted, it will demonstrate that the models used in this thesis provide a comprehensive explanation for adaptation at multiple cortical levels, accounting for both high-level and low-level adaptation and interactions between them.

6.4 Discussion and conclusion

This chapter constructed a multi-layered LISSOM network to replicate the main cartoon face results of the Xu et al. (2008) paper, which shows that low-level curve aftereffects can propagate to higher cortical areas and affect face emotion judgement. The modelling results showed that such a transfer of aftereffects can be achieved based on the same mechanism used to achieve face identity and face gender aftereffects.

This chapter extended the human experiment by computationally simulating more adaptation conditions, in order to see what a complete aftereffect curve can be like. Our previous modelling and experimental work on TAE and face gender all show an S-shaped curve, with a decline in aftereffect strength for test patterns sufficiently different from the adaptation stimulus. This chapter made predictions that a similar explanation should apply to curvature/emotion aftereffects as well. It strongly predicted an S-shape, i.e., that sufficiently large curvature values would lead to a lower effect. This

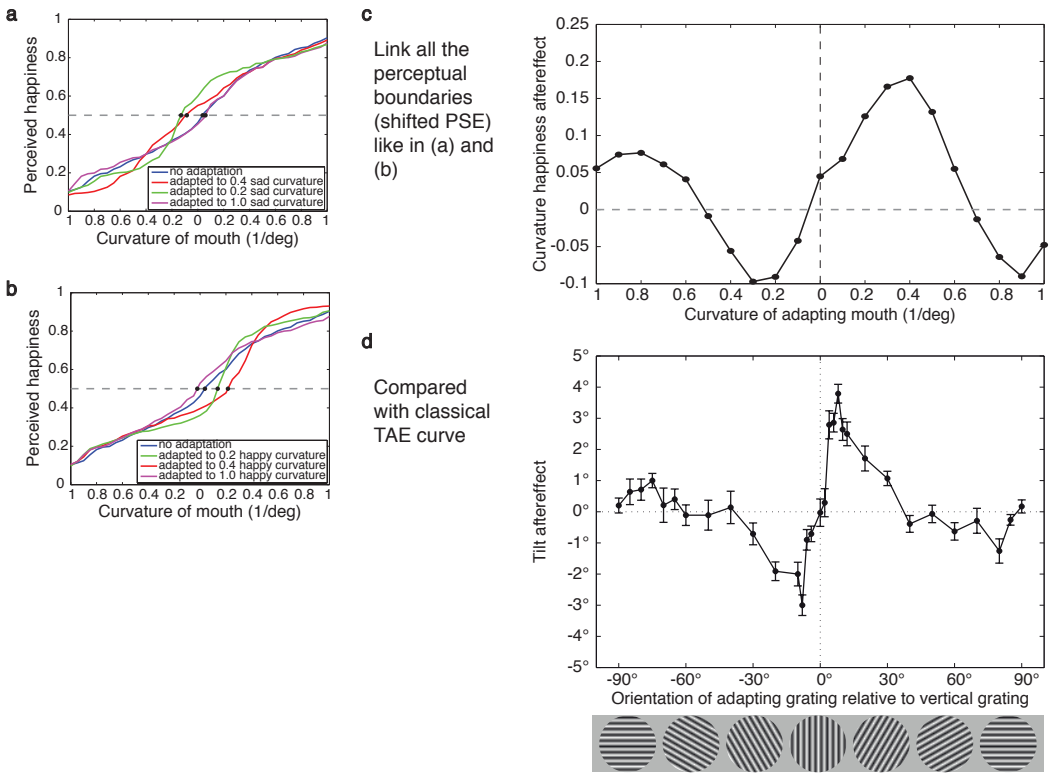


Figure 6.9: Prediction from the emotion aftereffect modelling and comparison with the classical TAE curve. **a**: Decoded perception on the test of the cartoon faces under the following conditions. No adaptation (baseline, blue); adaptation to the 0.2 (green), 0.4 (red) and 1.0 (magenta) sad curvature. For each condition, the decoded happiness was plotted as a function of the mouth curvature of the test faces. The points at happiness 0.5 (neutral emotion) are shown as solid black dots. **b**: Decoded perception from testing cartoon faces under the following conditions: No adaptation baseline (blue); adaptation to the 0.2 (green), 0.4 (red) and 1.0 (magenta) happy curvature. Otherwise the same as **a**. **c**: Prediction of the cartoon face emotion aftereffect curve if the full range of curves were tested. Twenty-one adaptation conditions like in **a** and **b** were tested and their happiness 0.5 (perceptual boundary) points were plotted as a function of the mouth curvature of the adapting mouths. **d**: The classical TAE curve from Mitchell and Muir (1976).

prediction can be tested in humans, potentially helping to constrain the properties of the neurons underlying the effect.

Note that this model focused on a subset of the results presented by Xu et al. (2008). They also conducted experiments where the test stimuli were photographic faces and showed that similar aftereffects could be obtained from adaptation to the real faces or to the same curved lines as above. These results further support the idea that low-level adaptation can be transferred to high-level cortical areas and can then lead to high-level aftereffects, and can be tested similarly in the model. However, to do so a significantly larger model will be needed, to ensure that V1 has a high enough density of units for responses to be reliable to natural images. The current size is sufficient for the simple cartoon faces used, but not to represent the multiple spatial scales (spatial frequency ranges) present in the natural face images. With a parallel implementation, it should soon be possible to simulate a large enough network to test the real-face conditions as well.

Chapter 7

Discussion, future directions and conclusion

This chapter summarises and concludes this thesis. The topics specific to each of the previous chapters have been discussed within each chapter, and here some general points that apply to all of the work presented in this thesis will be discussed first, followed by potential fields for future research. Finally, the topics this thesis has addressed (in terms of the problems proposed in Section 1.1) are summarised.

7.1 General discussions

In this section, some general points that apply to the modelling and experimenting on face aftereffects will be discussed.

7.1.1 Decoder and neural adaptation

In this thesis, the decoders used for simulating tilt (Section 3.2.4) face identity (Section 3.2.3), gender (Section 4.2.2) and emotion (Section 6.2.3) aftereffects have one common property — they are "unaware" of the neural response changes induced by adaptation, and they used a fixed strategy to decode the adapted responses. Therefore,

these decoders produced mismatched results which were interpreted as adaptation aftereffects, since the inducing stimuli are identical before and after adaptation. This approach was adopted because much psychological and physiological literature supports this simple type of model (Sutherland, 1961; Coltheart, 1971; Clifford et al., 2000; Jin et al., 2005; Schwartz et al., 2007), and theoretical work also suggests that the "unaware" approach is more consistent with behavioural data (Seriès et al., 2009). However, these studies are all based on low-level adaptation, and the real mechanisms by which perceptions are decoded and lead to behaviour, especially for high-level stimuli such as faces, still remain unclear.

Another important aspect of adaptation modelling in this thesis is that lateral only inhibitory connections are modified during adaptation simulation in the LISSOM-based models (Section 3.2.5, 4.2.2 and 6.2.3). As mentioned earlier, previous work showed that inhibitory updating was necessary and sufficient for LISSOM to duplicate the TAE (Bednar and Miikkulainen, 2000). Yet, the general mechanism underlying adaptation may not be that simple, especially since adaptation in high-level cortical areas is not that well understood. As indicated in Section 3.2.5, the underlying circuit is more complex than what was simulated in the model, and excitatory adaptation could also lead to similar aftereffects via disynaptic inhibition. Future work on elucidating the underlying mechanisms of visual adaptation will be needed to uncover the precise locations of adaptation, but this thesis predicts that such locations will be similar for both types of aftereffects.

7.1.2 Norm-based vs. exemplar-based modelling

The simplified model proposed in Section 4.3 is a type of "multichannel exemplar-based model". It showed that gender FAE reflects adaptation in neurons that are each tuned to a specific, though broad range of masculine or feminine faces (cf. Zhao and Chubb, 2001), while orientation selective neurons are tuned to a relatively narrow range of orientations. The experimental results for the FAE (see Section 5.2.5) are

consistent with this model.

To place these results in the context of other theories, it is important to consider previous ideas about how faces might be represented neurally in face space (Valentine, 1991). According to the face space theory, a particular face is a point in an underlying multidimensional space, where each dimension is a feature or attribute of the face (such as eye spacing or gender), and the centre of the space along all such dimensions is the mean face (see Section 2.4.1 for a review). Given this idea of face space, there has long been a debate about the two main types of neural representations of the space: norm-based and exemplar-based. The norm-based theory proposes that the mean face has a special norm status, and that the neural representations of other faces are based on deviations from this norm (see Rhodes et al., 1987 for the original work). More specifically, norm-based models typically employ two opponent pools of neurons selective for opposite stimuli in a dimension; with different attribute values represented as different balances between activations of these two pools (e.g., Rhodes and Jeffery, 2006). Multichannel exemplar-based theories instead employ a wide variety of neurons with different tuning curves, each most responsive to some particular face (i.e., combination of features), which suggests that faces are represented by activity in the neurons that respond most strongly to that specific combination of attribute values (Valentine, 1991; Valentine and Endo, 1992). More details about the comparison between these two kinds of theories can be found in Section 2.4.2.

The proposed FAE model in Section 4.3 is one example of a multichannel exemplar-based approach, albeit with broader tuning than one may have expected from low-level multichannel models. The important feature of this model is not the tuning width, but the bell shape of the tuning curve that makes each neuron respond best to a specific range of stimuli. The experimental results presented in this thesis support the idea that face-selective, or at least gender-selective neurons have tuning that limited in extent for any particular neuron. In addition, the population of such neurons is distributed across the range of face shapes, rather than occurring in two separate pools flanking the face

norm, as in norm-based opponent coding theories. Neurons preferring each possible stimulus have equivalent properties, and there is no special status in the model for a "norm" face, just as there was no special status for "vertical" in the orientation model. Instead, in this model, the importance of a norm lies in its role in defining a perceptual boundary in a specific perceptual task, whether for gender, distortion, identity, or other facial attributes. That is, the norm's role is during decoding the perceptual boundary (before or after adaptation), rather than to build the underlying neural representation.

Many studies have instead argued that a norm-based representation is necessary to explain high-level face aftereffects of identity and distortion. For example, Jeffery et al. (2010) recently argued that both children and adults exhibit norm-based face coding for face distortion, because they showed larger aftereffects for a large distortion value than for a smaller one. However, they only tested two widely separated data points, which does not provide enough information to distinguish between a linear increase and the S-shaped curve predicted by the multichannel models. More importantly, it is not clear how artificial distortion relates to processing for real faces. Gender is a natural face-space dimension and a clear source of variance in real populations, and thus presumably in neural representations, whereas distortion leading to non-human-like faces is not. The manipulated distortion may thus uncover different mechanisms than natural dimensions such as sex differences. Other studies reporting evidence for norm-based representations, such as Leopold et al. (2001), generally only examine a relatively limited range of values not far from the mean, and thus have not tested whether the aftereffect curve will decrease for higher values. They also focused on face identity rather than face gender, and it is possible that identity and gender of faces are encoded differently. However, physiological evidence by Leopold et al. (2006) showed that face-identity selective neurons in the monkey anterior inferotemporal cortex (AIT) region were tuned to a similar face space found in human psychophysical experiments (Leopold et al., 2001). The norm face had the lowest response, and the face-selective responses were shaped by the contrast between the test face and the face norm. This is

the only evidence so far that has implied specific neural implementation of norm-based encoding.

Of course, the "norm" stimuli are clearly special in some respects. For example, faces near the norm face are considered very attractive by human participants (Rhodes et al., 2003), and vertical orientations are more common in natural scenes because of their special status of balancing against gravity. However, whether the "specialness" of processing for these stimuli is "hard-coded" explicitly in the neural representation, or whether it emerges through general developmental and information-processing mechanisms, is an open question.

Other recent work has found evidence for the multichannel representations of relatively high-level processing of faces. Chen et al. (2010) very recently found a similar S-shaped curve for face viewpoint aftereffects, and Calder et al. (2008) have found evidence that face eye gaze aftereffects are better explained by a multichannel than a two-pool opponent approach. Both of these studies support the general idea of multichannel processing for visual stimuli, but both focus on dimensions that are closely related to orientation and direction, for which the S-shaped curve is already well established. It may be argued that the gender dimension used in this chapter is more clearly "high-level" than the gaze or viewpoint direction, and thus a better test of the hypothesis that high-level processing uses mechanisms similar to those established for low-level features like orientation.

The norm-based approach could be evaluated more rigorously if it were formulated as a computational model based on low-level measurements from neurons, as for the multi-channel models. One interesting example of such encoding is for image contrast, which in many models is represented using sigmoidal tuning curves for ON and OFF neurons (centered around a zero-contrast norm). It might be possible to estimate contrast from these two populations of neurons, as in the face norm theories, to validate a norm-based model using neural data. Such a model would provide a useful counterpoint to the multichannel models presented in this thesis, to make the differences

between the two approaches more concrete.

These results should prompt a re-examination of previous claims for the norm-based encoding of faces. The same methods as those used in Chapter 5 can be applied to any other type of stimulus that can be generated along a continuum that has a natural boundary, in order to see whether the results apply more generally for high-level perception.

7.1.3 Differences between low-level and high-level adaptation

As shown in chapters 3 and 4, the biggest difference between modelling the low-level TAE and the high-level FAE is the size of the receptive fields (LISSOM-based models) or the neuron tuning width (exemplar-based models). Apart from these structural differences, in the LISSOM-based models, the learning rates of all types of connections are different between TAE model and FAE models (e.g., see Section 3.2.1), where FAE models learn more slowly, perhaps because the differences between face stimuli are more subtle than the differences between different orientations. Different learning rates are used to avoid over-fitting of the model to the training data, compensating for the differences in learning speeds. Similar constraints may apply to biological systems, requiring different levels of plasticity to learn the different types of stimuli.

While these models can account for the current experimental results involving aftereffect curves, there are likely to be other factors underlying adaptation that may differ between high and low levels. For instance, high-level adaptation differs in properties such as discrimination and temporal dynamics (see Section 2.2 for a review) which were not addressed in this thesis. The difference in discrimination performance between low-level and high-level aftereffects is particularly intriguing (Rhodes et al., 2007; Oruç and Barton, 2009; Yang et al., 2010). In this thesis, the average time participants took to respond in FAE experiments was longer than for the TAE (see Section 5.3), and this temporal difference may reflect differences in the underlying processes.

Another temporal issue is the presentation time for test stimuli. As shown in Figure

5.2 and Figure 5.6, the test-stimulus presentation time for both TAE and FAE was 2 seconds, while in previous TAE studies durations of as short as 0.1 second have been used. In this study, the relatively long test period gave participants time to recognize each face; a shorter time would have sufficed only for the simpler orientation discrimination. The duration was kept fixed at 2 seconds so that both conditions would match most closely.

Although indirect effects have been reported previously for the TAE, and are evident in the modelling results in Figure 4.5, it is not clear whether there are any indirect effects in the experimental results (Figure 5.4). Thus the absence of an indirect effect for the FAE experimental results should be interpreted with caution – this method may simply not have much sensitivity for indirect effects (for adaptation stimuli far from the decision boundary).

In general, we are far from understanding the contribution of low-level factors to high-level adaptation effects. While the cross-cortical transfer of low-level aftereffects was covered in Chapter 6, the theory proposed in this thesis has not yet addressed the similar but more complicated face aftereffects containing more low-level factors, such as face viewpoint adaptation (Chen et al., 2010) or face contrast adaptation (Oruç and Barton, 2010). Additional studies more directly examining the differences between low-level and high-level adaptation will be crucial for understanding these issues.

7.2 Future directions

In this section, the potential fields for future research using modelling and experimentation will be discussed.

7.2.1 Modelling

As discussed in Section 6.4, the model described in Chapter 6 has not yet duplicated the real face test cases in the work conducted by Xu et al. (2008). This is model

will require a much higher model resolution and hence much more simulation time to be able to robustly handle the variation in the photographic faces. The next step for modelling is to extend the current model described in Section 6.2 to handle this variation by including many more neurons and connections.

As mentioned above, an important perceptual consequence of face adaptation — discrimination performance — was not addressed in this thesis. Previous work (e.g., Blakemore and Nachmias, 1971) has already indicated that aftereffects can be characterised by at least two factors experimentally: perceptual shift and detection thresholds. This thesis focused on the first factor and drew conclusions based on it. It will be interesting and important to investigate if similar approaches are used to determine how detection thresholds change in face space. This measurement is a proxy for the cognitive function of discrimination. Much previous literature has studied whether adaptation can facilitate discrimination, but the results have been controversial. For low-level aftereffects, it has generally been suggested that there is facilitation of discriminability around the adaptor (see e.g. Abbonizio, 2002), although the evidence for this is inconsistent. For face aftereffects, one study did not find any facilitation (Rhodes et al., 2007), but recent work claimed that small but significant facilitation can be measured (Oruç and Barton, 2009; Yang et al., 2010). Modelling work can provide subtle and precise prediction of discrimination change. It will be crucial to create a theory that can address the relationship between the change in perceptual shift and detection thresholds during adaptation, and establish a verifiable model to unify these two phenomena.

The temporal profile of face adaptation was also not addressed in this thesis. Though a longer time for making judgements was observed during the FAE experiment (see Section 5.3), this has not been considered in the model. It is also worth considering how to incorporate adaptation decay into the current model, in particular how to constrain the modelled dynamics to be consistent with the previous experimental results (Leopold et al., 2005; see Bednar and Miikkulainen, 2000 for modelling approaches).

7.2.2 Experiments

The next step in the experimental work is to conduct psychophysical experiments to verify the predictions made in Section 6.3. Note that as mentioned earlier, in the experiments conducted by Xu et al. (2008), the curvature of cartoon and real faces are their natural limitation. It is possible that the tuning bandwidth of "emotion" neurons is so wide that no "S-shaped" effect can be observed for the most curved adaptation condition that is feasible to test.

To further investigate the role of a "face norm", it will be interesting to see whether the S-shaped curve can be obtained by non-average face adaptation. In all FAE experiments described in this thesis, the morphing strength of adapting a face has been defined with respect to the average face. What kind of aftereffect curve can be obtained if the adapting and test stimuli are chosen from morphing trajectories that do not pass through the average face (e.g. a morphing trajectory between face A and face B through face C)? Experiments on adaptation to the average face based on the methods provided in Chapter 5 can also be conducted. Future experiments in these directions can help understand the real role of a "face norm".

Alternative experimental work can also employ cross-subject tests on both the TAE and the FAE, in order to quantify the relationship between their low-level and high-level tuning bandwidths. This thesis has found qualitatively similar patterns of the S-shaped curve across subjects. However, there is some individual variability, which it will be useful to analyse. For example, a participant who shows a broader curve in the FAE experiment could show a broad curve in the TAE experiment as well. So far, only two participants (CZ and JB) have taken part in both experiments, and thus no conclusions about this relationship can be drawn. Future work can focus on collecting matching TAE and FAE data from the same participants, so that individual differences can be analysed across the two experiments.

Another important issue is the face tuning bandwidth estimation. As mentioned in the thesis, the tuning bandwidth was broader for the FAE than what was expected

from the TAE results. However, rigorous comparison is not possible, because the full range of the face-related dimensions is unknown and thus how broad the tuning is in relation to the available range cannot be determined. On the other hand, by matching the S-shaped curves for the TAE and FAE, it is possible to make predictions for the tuning bandwidth of gender-selective neurons, which could then be tested in imaging or animal experiments.

As discussed in Section 5.3, the current experimental method is time-consuming and can be improved to draw more accurate results from the participants more quickly. For instance, future work can test adaptive Bayesian methods to limit the range of test stimuli and approach the participant's boundary more rapidly (e.g., Watt and Andrews, 1981; Kontsevich and Tyler, 1999).

7.2.3 Contribution of low-level factors on high-level aftereffects

This thesis revealed a potential similarity between low-level and high-level aftereffects, while opening up a broad range of possibilities for further investigating the complicated interplay between them, in particular the contribution of low-level factors on high-level aftereffects.

The work conducted by Xu et al. (2008) showed that low-level aftereffects can be transferred to high-level judgements, and some portions of the observed gender FAE may be due to low-level rather than high-level adaptation. One way to address this is to simultaneously adapt to low- and high-level stimuli, and to observe how the results differ from the transfer case. Apart from the curve adaptation, it is also worthwhile to include other low-level adaptors such as oriented gratings (as in the TAE) or illumination contrast (see Oruç and Barton, 2010). As a famous study target, emotional perception for the Mona Lisa can also be investigated in order to see how subtle low-level and high-level aftereffects can occur simultaneously.

Research on the complicated interplay between low- and high-level aftereffects should be conducted using both computational models and psychophysical experi-

ments, where both provide constraints to help us understand the underlying interactions. One may try to precisely predict, e.g., *A* amounts of low-level features will lead to *B* magnitude of high-level aftereffects.

7.2.4 Towards a combined multi-dimensional face adaptation framework

This thesis has provided preliminary theories for face identity, gender and emotion aftereffects, and suggested that similar mechanisms lead to both low-level and high-level aftereffects. With the experimental data constraining three dimensionalities, it is possible to combine these dimensionalities using a higher density of cortical sheets, more computational power, and well-tuned model parameters to provide a general theory for multi-dimensional face space. Ideally, one would model a three-dimensional or higher face space where the dynamics of any particular point (face) can be fully demonstrated and verified experimentally. Such a three-dimensional system should be a good start to gain an understanding of the complicated properties of visual aftereffects in general.

7.3 Conclusion

This thesis focuses on understanding the underlying neural mechanism of face adaptation from both a computational and psychophysical perspective. System-level computational models were first proposed, and then the results inspired later models and experiments on low-level and high-level aftereffects.

Chapter 2 reviewed the background topics covering the computational and experimental aspects of face adaptation.

Chapter 3 was motivated by previous experimental work on face identity aftereffects. This chapter tries to answer how face aftereffects can arise from neural activities and adaptation in a computational model. It established a comprehensive computational model and showed a perceptual shift representation for a two-dimensional face

space.

Chapter 4 was inspired by the results from Chapter 3 then focusing on a narrowed-down but more systematically addressed question: are face gender aftereffects similar to tilt aftereffects? This chapter built simplified computational models to show that similar aftereffect curves can be produced from a similar cortical mechanism.

Chapter 5 tested the similarity predicted in Chapter 4 using psychophysical experiments. It used an identical paradigm to test both the TAE and FAE, and showed results that were consistent with the prediction.

Chapter 6 tried to answer a further question of how low-level and face aftereffects transfer across low-level and high-level cortical areas. It used a similar modelling approach as in previous chapters to duplicate a cross-cortical transfer of face aftereffects, and again predicted an S-shaped curve that can be tested in experiments. Future modelling work can more fully address the transfer from low-level adaptation to photographic faces.

The novel contributions of this thesis consist of four main aspects: a model for the representation of multi-dimensional face space, both models and psychophysical experiments addressing the similarity between a gender FAE and the TAE, and a preliminary model investigating the transfer of face aftereffects from low- to high-level cortical areas. These results represent significant progress in understanding how face adaptation may be represented neurally, and what face processing shares with processing for simpler stimuli. This work suggests numerous additional computational and experimental studies that will help to understand interactions between low- and high-level adaptation.

Bibliography

Abbonizio, G. (2002). Contrast adaptation may enhance contrast discrimination. *Spatial Vision*, 16:45–58(14).

Addams, R. (1834). An account of a peculiar optical phenomenon seen after having looked at a moving body, etc. *London & Edinburgh Philosophical Magazine and Journal of Science*, 5:373–74.

Afraz, S.-R. and Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Res*, 48(1):42–54.

Aldrich, J. (1997). R. A. fisher and the making of maximum likelihood 1912-1922. *Statist. Sci.*, 12(3):162–176.

Antolik, J. (2011). *A Unified Developmental Model of Maps, Complex Cells and Surround Modulation in the Primary Visual Cortex*. PhD thesis, School of Informatics, The University of Edinburgh, Edinburgh, UK.

Bednar, J. A. (2002). *Learning to See: Genetic and Environmental Influences on Visual Development*. Technical report AI-TR-02-294, Department of Computer Sciences, The University of Texas at Austin, Austin, TX.

Bednar, J. A. and Miikkulainen, R. (2000). Tilt aftereffects in a self-organizing model of the primary visual cortex. *Neural Computation*, 12(7):1721–40.

Blakemore, C. and Nachmias, J. (1971). The orientation specificity of two visual aftereffects. *J Physiol (Lond)*, 213(1):157–74.

Blakemore, C. and Sutton, P. (1969). Size adaptation: A new aftereffect. *Science*, 166:245–7.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Booth, M. C. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex*, 8(6):510–23.

Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol*, 46(2):369–84.

Bukach, C. M., Gauthier, I., and Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. *Trends in Cognitive Sciences*, 10(4):159–66.

Calder, A. J., Jenkins, R., Cassel, A., and Clifford, C. W. G. (2008). Visual representation of eye gaze is coded by a nonopponent multichannel system. *J Exp Psychol Gen*, 137(2):244–61.

Chen, J., Yang, H., Wang, A., and Fang, F. (2010). Perceptual consequences of face viewpoint adaptation: face viewpoint aftereffect, changes of differential sensitivity to face view, and their relationship. *Journal of Vision*, 10(3):12.1–11.

Clifford, C. W., Wenderoth, P., and Spehar, B. (2000). A functional angle on some after-effects in cortical vision. *Proc Biol Sci*, 267(1454):1705–10.

Coltheart, M. (1971). Visual feature-analyzers and after-effects of tilt and curvature. *Psychological review*, 78(2):114–21.

Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 1st edition.

Dailey, M. N. and Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Netw*, 12(7-8):1053–1074.

Diamond, R. and Carey, S. (1986). Why faces are and are not special: an effect of expertise. *J Exp Psychol Gen*, 115(2):107–17.

Ekman, P. and Friesen, W. (1976). *Pictures of facial affect*. Consulting Psychologists Press.

Fang, F. and He, S. (2005). Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects. *Neuron*, 45(5):793–800.

Farah, M. J., Wilson, K. D., Drain, M., and Tanaka, J. N. (1998). What is "special" about face perception? *Psychol Rev*, 105(3):482–98.

Farivar, R., Blanke, O., and Chaudhuri, A. (2009). Dorsal-ventral integration in the recognition of motion-defined unfamiliar faces. *J Neurosci*, 29(16):5336–42.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200.

Gauthier, I. and Tarr, M. J. (1997). Becoming a "greeble" expert: exploring mechanisms for face recognition. *Vision Res*, 37(12):1673–82.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat Neurosci*, 2(6):568–73.

George, D. and Hawkins, J. (2005). A hierarchical bayesian model of invariant pattern recognition in the visual cortex. *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference*, 3:1812– 1817 vol. 3.

Gibson, J. and Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology*, 20:453–467.

Gilaie-Dotan, S. and Malach, R. (2007). Sub-exemplar shape tuning in human face-related areas. *Cereb Cortex*, 17(2):325–38.

Goldstein, A. and Chance, J. (1980). Memory for faces and schema theory. *Journal of Psychology*, 105:47–59.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci*, 15(1):20–5.

Grill-Spector, K., Sayres, R., and Ress, D. (2006). High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nat Neurosci*, 9(9):1177–85.

Gross, C. G. (1994). How inferior temporal cortex became a visual area. *Cereb Cortex*, 4(5):455–69.

Gross, C. G., Bender, D. B., and Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, 166(910):1303–6.

Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J Neurophysiol*, 35(1):96–111.

Hancock, P. J. (2000). Evolving faces from principal components. *Behav Res Methods Instrum Comput*, 32(2):327–33.

Hawkins, J. and Blakeslee, S. (2005). *On intelligence*. Henry Holt and Company, New York.

Haxby, J. V. (2006). Fine structure in representations of faces and objects. *Nat Neurosci*, 9(9):1084–6.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol (Lond)*, 195(1):215–43.

Jeffery, L., McKone, E., Haynes, R., and Firth, E. (2010). Four-to-six-year-old children use norm-based coding in face-space. *Journal of Vision*, 10(5).

Jin, D. Z., Dragoi, V., Sur, M., and Seung, H. S. (2005). Tilt aftereffect and adaptation-induced changes in orientation tuning in visual cortex. *J Neurophysiol*, 94(6):4038–50.

Johnson, M. and Morton, J. (1991). *Biology and cognitive development: the case of face recognition*. Cognitive development. B. Blackwell, Oxford.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–11.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol*, 71(3):856–67.

Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J Neurophysiol*, 80(1):324–30.

Köhler, W. and Wallach, H. (1944). Figural after-effects. an investigation of visual processes. *Proceedings of the American Philosophical Society*, 88(4):269–357.

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *J Neurophysiol*, 97(5):3155–64.

Konorski, J. (1967). *Integrative activity of the brain: an interdisciplinary approach*. University of Chicago Press, Chicago.

Kontsevich, L. L. and Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16):2729 – 2737.

Laming, D. (1997). *The measurement of sensation*. Oxford University Press, Oxford.

Leopold, D. A., Bondar, I. V., and Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–5.

Leopold, D. A., O'Toole, A. J., Vetter, T., and Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat Neurosci*, 4(1):89–94.

Leopold, D. A., Rhodes, G., Müller, K.-M., and Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proc Biol Sci*, 272(1566):897–904.

Little, A. C., Debruine, L. M., and Jones, B. C. (2005). Sex-contingent face aftereffects suggest distinct neural populations code male and female faces. *Proc Biol Sci*, 272(1578):2283–7.

Loffler, G., Yourganov, G., Wilkinson, F., and Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nat Neurosci*, 8(10):1386–90.

Logothetis, N. K. and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb Cortex*, 5(3):270–88.

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annu Rev Neurosci*, 19:577–621.

McKone, E., Kanwisher, N., and Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, 11(1):8–15.

Meytlis, M. and Sirovich, L. (2007). On the dimensionality of face space. *IEEE Trans Pattern Anal Mach Intell*, 29(7):1262–7.

Miikkulainen, R., Bednar, J. A., Choe, Y., and Sirosh, J. (2005). *Computational maps in the visual cortex*. Springer, Berlin.

Mishkin, M. and Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav Brain Res*, 6(1):57–77.

Mitchell, D. E. and Muir, D. W. (1976). Does the tilt after-effect occur in the oblique meridian? *Vision Research*, 16(6):609–13.

Mondloch, C., Lewis, T., and Budreau. . . , D. (1999). Face perception during early infancy. *Psychological Science*, 10:419–422.

Nordstrøm, M. M., Larsen, M., Sierakowski, J., and Stegmann, M. B. (2004). The IMM face database - an annotated dataset of 240 face images http://www2.imm.dtu.dk/ aam/datasets/datasets.html.

Oruç, I. and Barton, J. J. S. (2009). Improved face discrimination after face adaptation. *Journal of Vision [Abstract]*, 9(14):71.

Oruç, I. and Barton, J. J. S. (2010). A novel face aftereffect based on recognition contrast thresholds. *Vision Research*, pages 1–10.

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. *ICME 2005. IEEE International Conference on Multimedia and Expo*, page 5.

Park, J., Newman, L. I., and Polk, T. A. (2009). Face processing: the interplay of nature and nurture. *The Neuroscientist*, 15(5):445–9.

Peissig, J. J. and Tarr, M. J. (2007). Visual object recognition: do we know more now than we did 20 years ago? *Annu Rev Psychol*, 58:75–96.

Penton-Voak, I., Pound, N., Little, A., and Perrett, D. (2006). Personality judgments from natural and composite facial images: More evidence for a kernel of truth" in social perception. *Social Cognition*, 24(5):607–640.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333.

Perrett, D. I., Oram, M. W., and Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition*, 67(1-2):111–45.

Poulton, E. (1974). *Tracking skill and manual control*. Academic Press, New York.

Reddy, L. and Kanwisher, N. (2006). Coding of visual objects in the ventral stream. *Curr Opin Neurobiol*, 16(4):408–14.

Rhodes, G., Brennan, S., and Carey, S. (1987). Identification and ratings of caricatures: implications for mental representations of faces. *Cogn Psychol*, 19(4):473–97.

Rhodes, G. and Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, 46(18):2977–87.

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W. G., and Nakayama, K. (2003). Fitting the mind to the world: face adaptation and attractiveness aftereffects. *Psychological science*, 14(6):558–566.

Rhodes, G., Maloney, L. T., Turner, J., and Ewing, L. (2007). Adaptive face coding and discrimination around the average face. *Vision Research*, 47(7):974–89.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2(11):1019–25.

Robbins, R., McKone, E., and Edwards, M. (2007). Aftereffects for face attributes with different natural variability: adapter position effects and neural models. *Journal of experimental psychology Human perception and performance*, 33(3):570–92.

Rodman, H. R. (1994). Development of inferior temporal cortex in the monkey. *Cereb Cortex*, 4(5):484–98.

Rolls, E. and Milward, T. (2000). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures.

Rossion, B., Kung, C.-C., and Tarr, M. J. (2004). Visual expertise with nonface objects leads to competition with the early perceptual processing of faces in the human occipitotemporal cortex. *Proc Natl Acad Sci USA*, 101(40):14521–6.

Schwartz, O., Hsu, A., and Dayan, P. (2007). Space and time in visual context. *Nat Rev Neurosci*, 8(7):522–35.

Sergent, J., Ohta, S., and MacDonald, B. (1992). Functional neuroanatomy of face and object processing. a positron emission tomography study. *Brain*, 115 Pt 1:15–36.

Seriès, P., Stocker, A., and Simoncelli, E. (2009). Is the homunculus 'aware' of sensory adaptation? *Neural Computation*, 21(12):3271–304.

Snowden, R. J. (1998). Shifts in perceived position following adaptation to visual motion. *Curr Biol*, 8(24):1343–5.

Sohal, V. S. and Hasselmo, M. E. (2000). A model for experience-dependent changes in the responses of inferotemporal neurons. *Network*, 11(3):169–90.

Stringer, S., Perry, G., and Rolls. . . , E. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol Cybern*, 94(2):128–42.

Sutherland, N. (1954). Figural after-effects, retinal size, and apparent size. *The Quart. J. of Expt. Psych.*, 6(1):35–44.

Sutherland, N. (1961). Figural after-effects and apparent size. *The Quart. J. of Expt. Psych.*, 13(4):222–8.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci*, 19:109–39.

Thompson, P. and Burr, D. (2009). Visual aftereffects. *Curr Biol*, 19(1):R11–4.

Tiddeman, B., Burt, M., and Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *Computer Graphics and Applications, IEEE*, 21(5):42–50.

Treisman, A. (1996). The binding problem. *Curr Opin Neurobiol*, 6(2):171–8.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–4.

Ungerleider, L. G. and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr Opin Neurobiol*, 4(2):157–65.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A*, 43(2):161–204.

Valentine, T. and Bruce, V. (1986a). The effects of distinctiveness in recognising and classifying faces. *Perception*, 15(5):525–35.

Valentine, T. and Bruce, V. (1986b). Recognizing familiar faces: the role of distinctiveness and familiarity. *Can J Psychol*, 40(3):300–5.

Valentine, T. and Endo, M. (1992). Towards an exemplar model of face processing: the effects of race and distinctiveness. *Q J Exp Psychol A*, 44(4):671–703.

Vidyasagar, T. R. (1990). Pattern adaptation in cat visual cortex is a co-operative phenomenon. *Neuroscience*, 36(1):175–9.

Watt, R. and Andrews, D. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, 1(2):205–213.

Webster, M. A., Kaping, D., Mizokami, Y., and Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, 428(6982):557–61.

Whitney, D. and Cavanagh, P. (2003). Motion adaptation shifts apparent position without the motion aftereffect. *Percept Psychophys*, 65(7):1011–8.

Xu, H., Dayan, P., Lipkin, R. M., and Qian, N. (2008). Adaptation across the cortical hierarchy: low-level curve adaptation affects high-level facial-expression judgments. *Journal of Neuroscience*, 28(13):3374–83.

Yang, H., Shen, J., Chen, J., and Fang, F. (2010). Face adaptation improves gender discrimination. *Vision Research*, 51(1):105–10.

Zhao, L. and Chubb, C. (2001). The size-tuning of the face-distortion after-effect. *Vision Res*, 41(23):2979–94.